The Dissertation Committee for Dianne In-Hye Lou Certifies that this is the approved version of the following dissertation:

Strategies for Deciphering the Genome

Committee:

Sara L. Sawyer, Supervisor

William H. Press

Tanya T. Paull

Christopher S. Sullivan

Lauren I.R. Ehrlich

Kyle M. Miller

Strategies for Deciphering the Genome

by

Dianne In-Hye Lou, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2014

Dedication

I dedicate this work to my parents and my family. Thank you for supporting me during this long journey and for believing in me, every step along the way. I hope I can continue to make you proud!

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Sara Sawyer, for everything she has done for me in the past four years. You took a chance on me when no one else would and gave me the confidence to continue on this path. You believed in me even when I didn't believe in myself and most importantly, you never let me quit. Thank you for helping me realize my strengths and overcome my weaknesses. I honestly would not be in the position that I am in today if it wasn't for you. I hope I have made you proud as the first graduate of the Sawyer lab!

To my labmates and friends in the Sawyer lab: I really am going to miss you all. Susan, thank you for always making emergency orders for me, listening to me complain about anything and everything, and letting me convince you to go out to eat. You have always been the rock of the Sawyer lab and I can't imagine what it would have been like without you there. Nick (BigSexy725), my fellow senior graduate student, I already miss (and will continue to miss) our daily senior grad student meetings. Thank you for sharing your masterpiece, "The Saga of Mort", with me and doing your monkey dance on cue in front of complete strangers. Whenever I hear the words "monkey" or "taco" or "monkey poo", I will always think of you. To Maryska, my weekend running/torture and icecream/froyo/mochi/chipotle partner, thank you for never being judgmental, no matter how ridiculous I was being. Now that I think about it, you've rarely said no to me and for that, I am grateful. Alex, it's been nice to have a fellow MD/PhD student in the lab, especially since there aren't many of us around here. Thanks for all the bread, the laughter (many at your expense), and the random confidence boosts. And if I really do become a surgeon, I thank you for convincing me. Dr. Paul Rowley, we've been back-to-back in the office and at the bench for almost four years now. I can't say I'm going to

miss your sardines or the other smells you produce, but I will miss your always-cheerful demeanor and the way that shared equipment always ends up magically near or on your bench. Jeff, my sequencing team partner, it has been a pleasure working with you. Thank you for sharing your food with me (whether it was intentional or not) and for all the laughs along the way. Dr. Sandie Shan, although our time together was relatively short, you have definitely been a great addition to the lab. Thank you for all the favors you have done for me while I was frantically writing my last paper and my dissertation. Ross, Mimi, and Kim, my awesome and talented undergrads, I hope you have learned at least a little bit during your time with me. I am grateful for all your hard work and patience with me!! Dr. Ann Demogines, thank you for helping me through some rough patches during graduate school and giving me great advice when I needed it the most. Scott, thank you for singing out loud late at night in the lab when you thought no one else was there and the extra large jar of Nutella that I am still working my way through.

To my best friend Christien Kluwe, we've been through a lot over the last 8 years. Thank you for looking after my Bells during my all-nighters in the lab and for always being on my side. It would have been really difficult to go through this program without your companionship and I hope we will continue to stay friends over the years. Supriya Pai, my dear friend, you've always been a gracious host when I visit and your positive outlook on life is always an inspiration to me. To my Henderson 언니들, who were essentially my family during my years at Rutgers, thank you for all the food and Dunkin Donuts iced coffee and for my very first stethoscope! Hopefully, I will be able to rejoin you all soon! To everyone in the Ellington lab, past and present, thank you for taking me in as one of your own and letting me hang out in your lab!

I would also like to sincerely thank my committee members for all their advice and guidance. Dr. Bill Press, it's been a great pleasure and honor working with you on

the sequencing project. Your words of encouragement and wisdom have been very valuable to both Jeff and me. Dr. Chris Sullivan, thank you for always making time to meet with me and giving me great advice on the herpes project. Thank you for volunteering to be the first to tweet my future paper and for always being so enthusiastic about my work! Dr. Tanya Paull and Dr. Kyle Miller, I can't say enough how much of a blessing it has been to have two leaders in the field of DNA repair on my committee. Thank you for all your advice and for sharing your reagents and equipment with me. Dr. Lauren Ehrlich, your words of encouragement and expert FACS advice (which has revolutionized the way we do FACS in our lab!) have always been invaluable. Thank you!

Last, but certainly not least, I would to thank my parents for their endless support throughout this long training process. Although they don't quite understand why I have chosen this path (and sometimes, neither do I), they have always encouraged me to pursue my dreams and goals. They have taught me that hard work and patience perseveres in any situation and that I can achieve anything as long as I am determined to do so. I am so happy that I can finally give them a definite answer to the question "When are you graduating?" Mom, I know that I don't say this enough, but thank you for all that you have sacrificed and done for me. I can't even begin to imagine how hard it was for you to raise my brother and me as a single mother, all the while running a business full-time. I am so glad that you have found your soulmate and I hope that the two of you will continue to live a long and healthy life together. Dad, thank you for always letting me know how proud you are of me and treating me like I am your own (and also for supplementing my poor graduate student salary when my mom isn't watching). I am forever indebted to you for making my mom happier than she has ever been. 엄마, 아빠, 사랑해요!!

Strategies for Deciphering the Genome

Dianne In-Hye Lou, Ph.D.

The University of Texas at Austin, 2014

Supervisor: Sara L. Sawyer

The development of highly sophisticated technologies has ushered in the era of the genome. Most importantly, high-throughput sequencing technologies has vastly expanded the number of available genome projects of many different organisms. One of challenges that we now face is in understanding the information encoded within these genomes. Within each chapter of this dissertation, information from existing genome projects are used to answer fundamental biological questions related to human disease and an attempt to further advance new technologies is made. In chapter 2, I describe a novel method that decreases the error rates associated with next-generation sequencing technologies, allowing for the investigation of more complex and heterogenous samples relevant to many biological systems. In chapter 3, I use available primate genome projects to understand the evolutionary trajectory of two DNA repair genes, whose defect increases the development of breast and ovarian cancers. Finally, in chapter 4, wild-type primate alleles are used as tools to uncover novel mechanisms in the lifecycle of viruses. Although seemingly non-overlapping, each of these studies is centered around using the sea of information that is now readily available in order to decipher the many secrets encoded by genome.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to Dissertation

I have been blessed with a unique opportunity to work on several very different projects during my graduate studies. As such, each project deserves its own introduction and is therefore included at the start of every chapter. In the place of a formal introductory chapter, extended abstracts for each project are presented below.

CIRCLE-SEQUENCING: A NEW ERROR-CORRECTING METHOD FOR HIGH-THROUGHPUT SEQUENCING DATA

Unlike traditional Sanger sequencing machines that can only determine several hundred sequences at a time, high-throughput sequencing technologies are able to simultaneously sequence millions of different DNA molecules from complex mixtures. These machines have revolutionized the study of biological systems by offering vastly increased throughput and by dramatically reducing costs (1, 2). For example, if one were to sequence all 3 billion bases of the human genome, it would take at least 3.75 million reactions and cost $7.2 million using Sanger sequencing. With next-generation sequencing, one run can potentially sequence the entire human genome for only $6000 (3). Therefore, it is no wonder that scientists have heavily utilized these technologies in the recent years to address outstanding research problems and to ask provocative new questions.

However, a fundamental problem with all next generation sequencing technologies is that there is a relatively high rate of incorrectly identified DNA bases (typically on the order of 1%) (1). This problem can partially be overcome in cases where a homogenous sample, such as the genomic DNA of one individual, is sequenced by simply reading every base at least 30 times. By doing so, several errors generated by the

machine can be identified and eliminated. Unfortunately, this solution is not completely effective at eliminating sequencing-generated errors and is not applicable for heterogeneous samples, like tumors, where one wishes to actually identify genetic variants. For this reason, the error rate problem renders these new technologies unable to address many of the important problems in biomedical research.

Several attempts have been made to address the error rate problem. These methods primarily involve covalently attaching a unique identifying sequence, also referred to as barcodes, to individual DNA molecules (4-8). The barcoded samples are then amplified using a DNA polymerase to generate multiple copies of each sequence. By doing so, each molecule of DNA can be read several times and a consensus sequence can be derived. While these methods can achieve impressively low error rates, this is an extremely inefficient process. First, the ligation of barcodes to the DNA molecules is very inefficient and this excludes a majority of the sample from being sequenced. In addition, the number of copies that are generated during PCR is extremely difficult to control, often times leading to under- or over-amplification of the sample. Lastly, the cost associated with optimizing the conditions described above can be inhibitory, especially when sequencing large genomes.

To circumvent these issues, we have developed a unique library preparation strategy and accompanying computational correction scheme called circle sequencing, described in Chapter 2. In this method, copies of individual DNA templates are circularized and copied multiple times in tandem with a rolling circle polymerase, resulting in physically linking of each copy. These products are then sequenced on any high-throughput sequencing machine. Each read produced is computationally processed to obtain a consensus sequence of all linked copies of the original molecule. Because the

circle-sequencing protocol precedes standard library preparations, it is suitable for a broad range of sequencing applications.

RAPID EVOLUTION OF THE DNA REPAIR GENES *BRCA1* AND *BRCA2*

The maintenance of chromosomal integrity is an essential task of every living organism and cellular repair mechanisms exist to guard against insults to DNA. Because the genome encodes all the information needed to sustain life, it is imperative that DNA repair mechanisms exist to recognize and fix damages to the DNA. This is exemplified by the fact that mutations in genes that are involved in DNA repair often lead to devastating diseases or are associated with an increased risk for cancer (9).

The most detrimental type of DNA damage is the double strand break (DSB). Immediately after the induction of a DSB, an elaborate cascade of signaling events occurs, resulting in the initiation of the DNA damage response (9). The DNA damage response primes the nuclear environment in preparation for the repair of these lethal lesions by activating the appropriate signaling molecules, recruiting essential DNA repair factors, and halting the progression of the cell cycle to allow time for efficient repair. In humans, two major repair mechanisms are devoted to the repair of DSBs: the non-homologous end-joining (NHEJ) and the homologous recombination (HR) pathways (10). Classically, the NHEJ pathway is considered to be an error-prone process in which the two ends of the broken chromosome can undergo resection and direct ligation, resulting in mutagenic deletions or insertions at the site of damage. This process of DSB repair predominates during most phases of the cell cycle because it does not require a sister chromatid to serve as a template for repair. The homologous recombination

pathway, on the other hand, utilizes a template for perfect repair and is most active during the S and G2 phases of the cell cycle.

Given the importance of these processes, it is expected that proteins involved in the DNA damage response and repair pathways would be evolutionarily conserved, exhibiting very minimal sequence change over time. However, several DNA repair proteins have been shown to be evolving rapidly, accumulating mutations at a rate that is much higher than expected by chance (11, 12). In particular, *BRCA1*, an essential gene whose protein product plays several roles in DDR and repair, exhibits a striking signature of rapid evolution that is characteristic of a gene under intense selective pressure (12-17). Although these studies have been enlightening, sequences from very diverse mammals were utilized. This can be especially problematic when identifying specific residues undergoing rapid evolution because different selective forces may have uniquely acted upon different mammalian species. To better understand the pressures that the *BRCA1* gene has faced in humans and non-human primates, an evolutionary analysis using a more extensive dataset that includes a number of closely related primate species is ideal.

In chapter 3, we explore the evolutionary history of *BRCA1* and *BRCA2*, another hereditary determinant of breast and ovarian cancers, in primates. We show that specific residues in the primate BRCA1 and BRCA2 proteins are evolving rapidly and that these residues are different than the ones previously reported in the literature. We also find considerable variation within primate populations and provide evidence for recent selection in chimpanzee populations. We put forth a hypothesis in which pathogens may be driving the evolution of these essential genes.

THE DNA REPAIR PROTEIN NBS1 IS A BARRIER TO HSV-1 REPLICATION

On October 22nd, 1932, Dr. W.B suffered a bite from a rhesus macaque during the course of an experiment. He treated his own wound with simple first aid measures and continued on with his study. A few days later, the superficial lesion became noticeably red and swollen, eliciting pain in the patient. He was admitted to the hospital shortly thereafter, with a low-grade fever and swelling of the affected arm. As his condition began to improve, small fluid-filled vesicles began to form at the site of penetration. However, several days after the incident, the patient began to show signs of rapid neurological deterioration and eventually succumbed to his death (18, 19).

The pathological agent responsible for the patient's demise was found to be the macaque simplex virus 1 (MHV-1), also known as herpes B virus. Much like herpes simplex virus 1 (HSV-1) of humans, MHV-1 is a ubiquitous virus in the macaque population that establishes a lifelong infection with periodic reactivation (20). Active viral shedding is usually asymptomatic in macaques, but can result in the same types of oral and ocular lesions induced by HSV-1 in humans. So how is it that MHV-1 causes such a devastating disease in humans while showing no signs of presence in its original host? The answer lies within the genome of the virus and the organism that it infects.

Viruses encode for a number of proteins that facilitate replication within a cell, many of which establish interactions with host proteins that further enhance the viral lifecycle. In cases where virus-host interactions are sub-optimal, viral replication may be considerably crippled or even non-existent. Conversely, when these interactions are too strong, uncontrolled amplification of the virus may cause serious harm to the host, resulting in death. Both of these scenarios represent a dead-end for the virus, resulting in a barrier to successful and sustained cross-species transmission of the virus. Therefore, it is within the best interest of the virus to accumulate advantageous mutations and fine-

tune these interactions to strike a balance these two outcomes, a process called viral adaptation. Even after the virus has adapted to its host, the virus may continue to evolve in an effort to be maintained within the host, in a similar dynamic to that described above called codivergence.

Genetic similarities between a host and a closely related non-host species can facilitate the introduction of a virus to a new species. In fact, the human population is constantly plagued by the emergence of new viral threats resulting from transmission events that cross species boundaries (21, 22). One classical example is that of the human immunodeficiency viruses (HIV), a result of simian virus transmission from primates to humans. Although simian immunodeficiency viruses (SIVs) cause little to no disease in their natural primate hosts, HIV infection in humans results in a progressive decline in CD4+ T-cells that eventually leads to immunodeficiency. This disparity in clinical outcomes could be a result of inadequate adaptation of HIV in the genetically similar, but not identical, human host. This newly forged relationship between virus and host has yet to strike the balance seen in codivergence.

As such, it is of considerable interest to understand the complex network of host-virus interactions so that one day, we may be able to predict the potential virulence of "new" viruses that threaten the human population (23). In chapter 4, we investigate the dynamic interaction between HSV-1 and the host DNA repair protein, Nbs1. We show that HSV-1 hijacks Nbs1 functions in a species-specific manner in order to augment the viral lifecycle. We identify a primate Nbs1 allele that is resistant to recruitment by HSV-1 and show that this variant may pose a significant barrier to infection by the virus. Further studies are warranted to determine if other primate simplex viruses are able to utilize the human Nbs1 protein for advancing the viral lifecycle. These types of analyses should be extended to other cellular factors that can potentially restrict or aid in the viral

lifecycle of herpes simplex viruses so that we can gain a better understanding of this particular host-virus interaction landscape. This will allow us to predict the pathogenic potential of these viruses if and when they gain a foothold in the human population.

Chapter 2: High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing

INTRODUCTION

High-throughput DNA sequencing has emerged as a revolutionary force in the study of biological systems. However, a fundamental limitation of these technologies is the high rate of incorrectly identified DNA bases in the data produced (1, 2). For instance, reports in the literature suggest that Illumina sequencing machines produce errors at a rate of ~0.1–1 x 10^{-2} per base sequenced, depending on the data-filtering scheme used (1, 24). These technologies typically produce billions of base calls per experiment, translating to millions of errors. When sequencing a genetically homogenous sample, the effects of erroneous base calls can be largely mitigated by establishing a consensus sequence from high-coverage sequencing reads. However, even high coverage does not eliminate all errors, and attempted verification of detected variants has often revealed the vast majority to be sequencing errors (for example, see refs. (25, 26)). Furthermore, the depth of coverage required for consensus building remains cost-prohibitive for large genomes such as the human genome. As a result, most human studies involving high-throughput sequencing have been limited to only a small fraction of the genetic information, such as the transcriptome, mitochondrial DNA, or a single chromosome. In contexts where rare genetic variants are sought, this error-rate problem presents an even more profound barrier. Examples of rare variant problems include the analysis of mutations in genetically heterogeneous tumors, identification of drug-resistance mutations in microbial populations, and characterizations in immunogenetics (such as B- and T-cell profiling). The mutations of interest in these types of samples may

be present at low frequencies, potentially even lower than the sequencing error rate itself. Here, the problem cannot be overcome with high sequence coverage because a consensus sequence of the heterogeneous sample will mask all variants.

To address this error rate problem, several closely related library preparation protocols have recently been described (4-8). A general schematic for these "barcoding" strategies is shown in Figure 2-1. Each individual DNA molecule in the input material is marked by the ligation of a uniquely identifiable sequence, or barcode (step 1). Barcoded products are then amplified by PCR (step 2), and the amplified pool is sequenced (step 3). Barcode identity is then used to computationally organize sequencing reads into "read families," where each read family consists of all downstream derivatives of a single starting molecule (step 4). A consensus sequence is then derived from the reads in each family, with a typical criterion being that the read family must contain at least three members before a consensus sequence is derived (4, 7).

Although barcoding strategies successfully lower the sequencing error rate, these methods have both theoretical and practical limitations that affect the accuracy and cost with which consensus sequences can be produced. First, the members of a read family are not independent copies of the original molecule. Errors that arise during the early stages of PCR, known as jackpot mutations, are amplified exponentially and can appear multiple times in a read family. Second, some templates may be amplified more or less efficiently due to differences in either the barcodes or the target sequences themselves (4). This bias, along with unavoidable variance in the sampling process, results in many read families being much larger or smaller than necessary, reducing efficiency. Third, if identical barcode sequences are ligated to multiple input molecules similar in sequence, incorrect assembly of read families can occur. This issue is especially problematic when sequencing highly similar molecules such as in amplicon libraries. Finally, sequencing

errors introduced into barcodes themselves contribute to inefficient formation of read families.

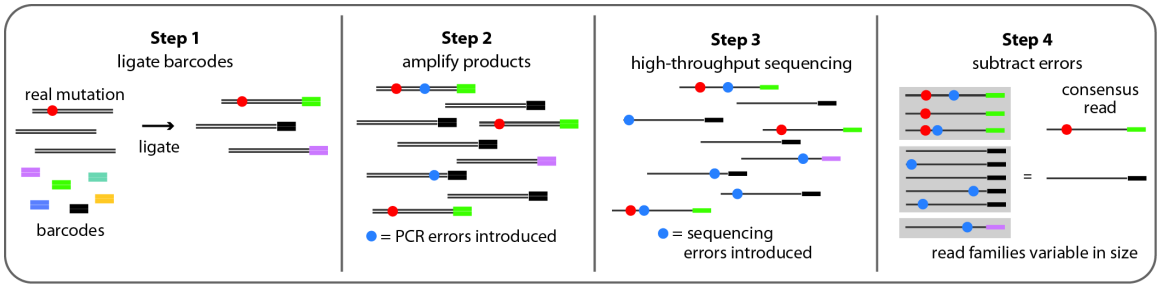


Figure 2-1: Overview of traditional barcoding methods. In traditional barcoding methods, adapters containing randomized nucleotide regions (barcodes) are ligated to each molecule in the DNA sample (step 1). The library is then amplified by PCR (step 2). Products are sequenced on the high-throughput sequencing platform of choice (step 3). Individual reads containing the same barcode are grouped into read families (gray boxes), and consensus sequences are derived (step 4). Errors generated during PCR amplification (step 2, blue circles) and during the sequencing process (step 3, blue circles) are removed bioinformatically.

In theory, all of these problems could be avoided if each read family were packaged and delivered as a single molecule, bypassing the need for barcodes to construct read families. Continued advances in the read lengths of major sequencing platforms have made such an approach possible. We have developed a unique library preparation method that (i) eliminates the use of barcodes, (ii) eliminates the effects of jackpot mutations by amplifying DNA templates in a way that does not propagate errors within read families, and (iii) physically links the repeated information comprising each read family so that it comes out of the sequencing process in the optimal proportions needed for efficient error correction. We show that our method produces high-throughput sequencing data with errors at only 8–10 x 10^{-6} of base positions sequenced and has an efficiency that is vastly improved over existing barcoding schemes. Our

library preparation method, called "circle sequencing," fits into existing high-throughput sequencing workflows, making it immediately available for a broad range of applications.

MATERIALS AND METHODS

Circle sequencing

Genomic DNA was extracted from the *S. cerevesiae* strain S288C, sheared, run on a 1.5% low-melting-point agarose gel, and a narrow slice corresponding to 150 bp was extracted. DNA was phosphorylated and denatured. 300 ng DNA (~3 pmol) was circularized per 20 μL reaction using CircLigase II ssDNA ligase (EpiCentre), and uncircularized DNA was removed with exonuclease. Exonuclease-resistant random primers and varying amounts of DNA circles were annealed and added to the rolling circle reaction consisting of reaction buffer, dNTPs, BSA, inorganic pyrophosphatase, uracil-DNA glycosylase, formamidopyrimidine-DNA glycosylase, and Phi29 DNA polymerase. For a detailed protocol, refer to Appendix A. Barcoded samples were prepared as in ref. (7).

Bioinformatic Processing

Our computational pipeline processes circle-sequencing data generated by paired-end reads. The structure of the tandem copies within each read pair is determined by detecting periodicity in each sequence and by aligning the pair of sequences to each other. A consensus sequence is then derived from the copies produced in combination with the base quality scores assigned to each. The junction of circularization in each consensus sequence is identified by performing a rotation-insensitive mapping of the consensus sequence to a reference genome. See ref. (27) for a detailed description.

RESULTS

Circle Sequencing: Library Preparation

Our library-preparation method, circle sequencing, is illustrated in Figure 2-2. The input material for this protocol can be chromosomal DNA, cDNA, amplicons, or any other DNA. The material is size-selected (through amplicon design, shearing followed by gel purification, etc.) such that the size of each fragment averages around 1/3 the anticipated read length from the high-throughput sequencing machine being used. Double-stranded DNA fragments are denatured and the resulting single-stranded DNA is circularized (step 1). Non-circularized products are eliminated by exonuclease digestion. Random primers are then annealed to the single-stranded circular DNA, and amplification is performed using the Phi29 polymerase. This polymerase possesses single-strand displacement activity that allows it to replicate continuously around the ligated circle, referred to as rolling circle amplification (step 2). The random primers also anneal to the newly synthesized single-stranded product and allow it to be converted into double-stranded DNA. The resulting double-stranded DNA products (step 2, lower) are concatamers consisting of multiple tandem copies (brackets) of the information in the original fragment. These products are sequenced (step 3), and the information in the tandem copies is used to form a read family and a consensus sequence (step 4). Any genetic variant that existed in the input material (red circle) will be present in all tandem copies whereas errors introduced by the Phi29 polymerase (blue circles in step 2) or by the sequencing process (blue circles in step 3) will occur independently and randomly throughout the template.

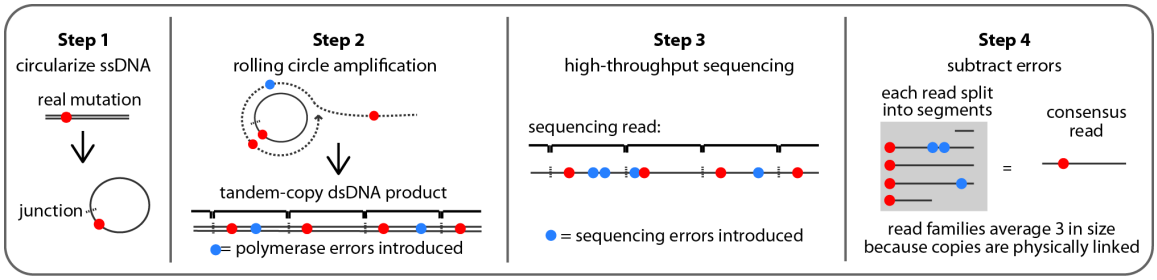


Figure 2-2: Overview of circle sequencing. In circle sequencing, DNA is denatured and single-stranded DNA is circularized (step 1). Random primers are annealed to circles, and Phi29 polymerase is used to perform rolling circle replication (step 2). This polymerase has strand-displacement activity so products contain tandemly linked copies of the information in the circle. Random primers and Phi29 polymerase turn long single-stranded copies into double-stranded DNA (step 2, lower). The tandem copies of information are sequenced using any high-throughput sequencing technology (step 3). Here, a single long read is shown for simplicity although paired-end reads were used in this study. Each read (or paired-end read pair) is then computationally split into the individual copies of the original circle, grouped into a read family (gray box), and used to generate a consensus sequence (step 4).

The rolling circle products generated in circle sequencing can theoretically be sequenced on any high-throughput sequencing platform that offers read lengths long enough to observe multiple repeats within the same product. Illumina technologies currently offer the highest throughput and cost-efficiency, with read lengths of up to 500 bases possible on the MiSeq platform using 2 x 250 paired-end reads. Our bioinformatic pipeline processes circle-sequencing data generated by paired-end reads. This pipeline identifies the repeating units of the original information in each read pair. It then uses these repeats, combined with base call quality scores, to derive a consensus sequence along with a consensus quality score for each consensus base. The pipeline also maps these consensus sequences to a reference genome.

One advantage of circle sequencing is that it is largely resistant to the effects of jackpot mutations that can occur in PCR. Errors will also be made during rolling circle amplification, but will not propagate within a read family because each linked copy is in-

dependently derived from the original molecule (Figure 2-3). An upstream PCR amplification step may be required for some applications of circle sequencing (e.g., for cDNA or amplicon libraries). Circle sequencing will not be able to mitigate the effects of jackpot mutations accumulated before templates are circularized. In such cases, care should be taken to minimize amplification cycles upstream of the circle-sequencing pipeline. Alternately, circle sequencing can also be applied directly to RNA templates (28).

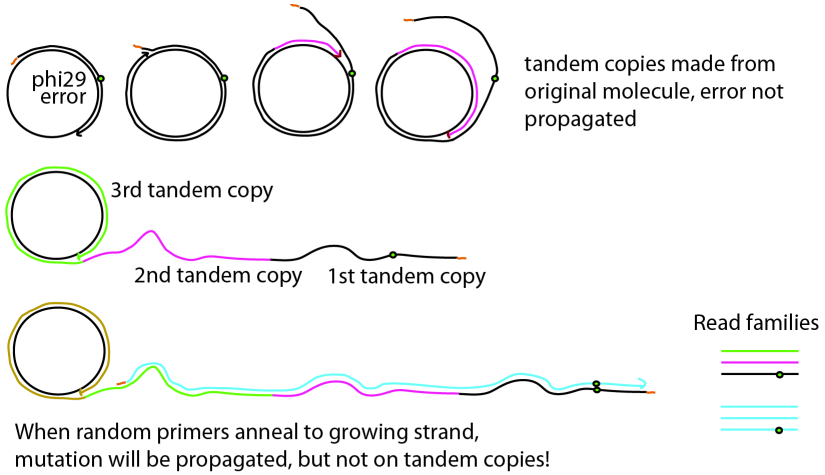


Figure 2-3: Error propagation in read families. In circle sequencing, Phi29 polymerase primes off of a random primer (orange) to generate a single copy of the circular template. After the polymerase completes the first copy (black), it displaces that strand and begins a second copy (shown in pink), then a third copy (shown in green), and so on. Errors incorporated by the polymerase during the replication of the first repeat (green dot) will only be contained within the first repeat (black) and not in subsequent repeats (pink, green, yellow).

Error Rate of Circle Sequencing

To measure the error rate of this method, we sequenced the ~12-megabase *Saccharomyces cerevisiae* genome. First, we used standard Illumina MiSeq sequencing to obtain 51X coverage of a haploid S288C strain. We identified 514 positions at which

there is strong evidence that the genome sequence of this strain differs from the published reference S288C sequence. Bases mapping to these questionable sites, or to repetitive sequences, were subsequently ignored throughout this study. Next, we sequenced the strain with circle sequencing and mapped the resulting consensus sequences to the reference genome. Error rates were calculated as the fraction of consensus bases that differed from the reference sequence. As a proof of concept that the circle-sequencing process is capable of eliminating sequencing errors, we calculated the error rates of consensus sequences formed by incrementally incorporating each repeat contained in each read pair. High-quality bases in the first repeat of each sequencing read had an error rate of 5.8×10^{-4} (Figure 2-4A). As expected, this error rate fell as the tandem repeats were used to correct the error in the first repeat. However, the effect was surprisingly small. The inclusion of subsequent tandem copies in our reads reduced this error rate only to 2.9×10^{-4} with two repeats and 2.7×10^{-4} with three repeats (Figure 2-4A). Because the circularized fragments used in the circle-sequencing pipeline are size-selected, but still vary in size according to a distribution, we recover four and sometimes more repeat units in some of the read pairs. The addition of subsequent repeats beyond three did not lower the error rate further, and the asymptotic value of the error rate achieved (2.8×10^{-4}) was not as small as would be implied by the informational redundancy obtained.

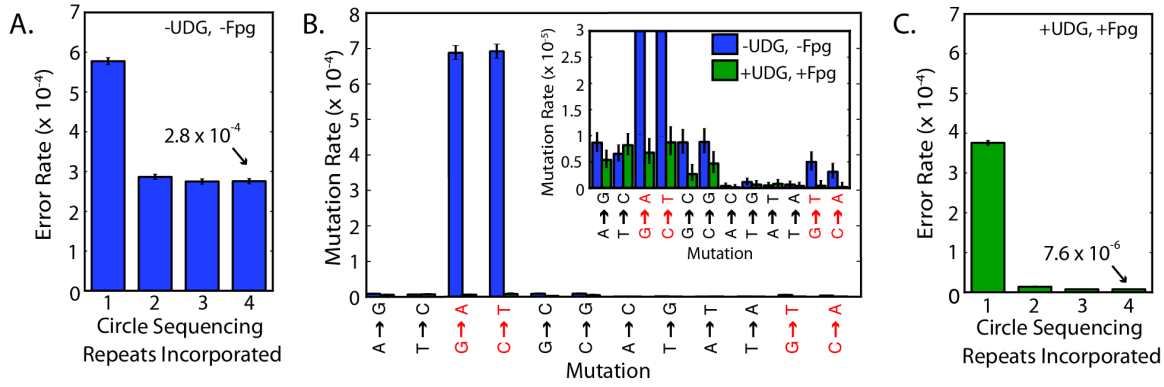


Figure 2-4: Circle sequencing decreases the error rate of high-throughput sequencing. (**A**) Shown is the error rate of circle sequencing upon incorporation of additional copies of the tandemly duplicated information. When just the first repeat is considered, only bases with quality scores greater than or equal to 30 are used. (**B**) The profile of the types of errors in circle sequencing consensus sequences (blue bars) reveals a striking signature dominated by G-to-A and C-to-T errors, consistent with base damage due to cytosine deamination. The addition of uracil-DNA glycosylase and formamidopyrimidine-DNA glycosylase during rolling circle amplification (green bars) dramatically eliminates the majority of errors (red) caused by DNA damage of the types targeted by these enzymes. The lower range is shown in more detail in the Inset. (**C**) As in **A**, the graph shows the error rate of circle sequencing upon incorporation of additional copies of the tandemly duplicated information, after the protocol was modified to include DNA repair enzymes.

One barrier to achieving lower error rates using circle sequencing could be DNA damage incurred by sequencing templates during the library preparation protocol. This damage would be especially problematic because, unlike random mutations introduced by the polymerase, a damaged base within the original circular template might be paired with the same incorrect base each time the polymerase replicates around the circle. This process would lead to the propagation of an error in all tandem copies of the circular information. These errors would then be seen as high-confidence base calls and effectively increase the overall error rate. In fact, DNA damage has been identified as a major source of error in standard barcoding-based approaches (7). To examine this possibility further, we analyzed the different types of erroneous base calls produced in

circle-sequencing consensus sequences (Figure 2-4B). Interestingly, we discovered that a large proportion of mismatches between the consensus sequences and the reference genome were G-to-A and C-to-T mutations. These types of mutations occur when a cytosine base undergoes spontaneous deamination to form uracil. Adenine is incorporated opposite of the uracil during synthesis of the complementary strand, propagating G-to-A and C-to-T transitions. We did not detect a substantial number of G-to-T and C-to-A mismatches indicative of oxidized guanine bases (8-oxo-guanine), the other common type of damage found to affect barcoded samples (7).

Deaminated cytosine and 8-oxo-guanine bases can be excised using the commercially available enzymes uracil-DNA glycosylase (UDG) and formamidopyrimidine-DNA glycosylase (Fpg). To test whether these specific types of damaged bases negatively affect the error rate of our method, these enzymes were included during the rolling circle amplification step. As shown in Figure 2-4B, their addition almost completely eliminated damage-induced errors (green bars). We speculate that circular templates that undergo the removal of these damaged bases are precluded from serving as substrates for Phi29 polymerase. We found that treatment of genomic DNA with UDG and Fpg before proceeding with conventional MiSeq library preparation resulted in no change in the mutation profile, suggesting that this damage to DNA is actually incurred during the circle-sequencing library preparation. After modifying our protocol to include these repair enzymes, we reexamined the impact of analyzing one, two, three, and four repeats in consensus building (Figure 2-4C). The inclusion of up to four repeats substantially improved the overall error rate from 2.8×10^{-4} (without enzymes) to 7.6×10^{-6} (with enzymes). Thus, the error-correcting power of circle sequencing is clear, but care must be taken to address damaged bases that arise during the preparation of these special libraries. It is interesting to note that the extent of DNA

damage present in DNA mixtures, and the resultant calling of erroneous bases during downstream sequencing, is only now becoming evident due to the extremely low error rates being achieved by our and other sequencing methods (7). We anticipate that accurate, high-throughput sequencing will provide increased resolution into many types of biological phenomena.

Efficiency of Circle Sequencing and Barcoding

An important metric to consider when selecting an error-correction scheme (i.e., barcoding versus circle sequencing) is cost. Cost is directly related to the efficiency of these methods in turning low-quality data into high-quality data. A major determinant of the overall amount of high-quality data produced by a method is how efficiently the method distributes raw sequencing data across all of the read families produced. The existence of read families that do not contain enough members to produce a consensus, or read families that contain substantially more members than necessary, represents wasted sequencing resources. To analyze this aspect, we define efficiency as the ratio of consensus bases produced to the total number of bases used to produce them. As described above, read families must have at least three members to build consensus sequences. If all read families consist of exactly three members, a perfect efficiency of 33% would be achieved. For circle sequencing, the size of read families is dictated by the lengths of the circularized molecules. To achieve perfect efficiency of 33%, input molecules must be exactly 1/3 the read length. However, any practical size-selection scheme will produce molecules with a distribution of sizes around this desired length. This distribution, and the use of paired end reads (discussed later in this section), results in the actual achieved efficiency being slightly lower than the ideal. In agreement with

this reasoning, we achieved an efficiency of 20.2% for circle sequencing (Figure 2-5A). One consensus base is produced for every five bases used to build read families.

For comparison, we calculated the efficiency of consensus sequence formation with barcoding using a dataset from a previously published study by Schmitt *et al*. (7). This barcoded dataset, derived from the M13mp2 phage genome, produced consensus sequences with an efficiency of 3.0% (Figure 2-5A). One consensus base is produced for every 33 bases analyzed. This efficiency is similar to previously reported barcoding efficiencies, which range from 1–8% (4, 5, 7). In this particular dataset, the authors used a sophisticated barcoding scheme called duplex barcoding. Here, the forward and reverse strands of each double- stranded input molecule are asymmetrically labeled with barcodes, allowing for the acquisition of either standard barcoding read families or, alternately, more elaborate read families consisting of at least three reads from each strand (i.e., at least six reads total). As would be expected because of the heightened read family criteria, duplex-barcoding consensus sequences were formed with an efficiency of only 0.8% with this dataset, substantially less than the efficiencies of either circle sequencing or standard barcoding.

For barcoding-based approaches, efficiency is dictated by the ratio of barcoded input molecules to total reads produced. If there are too many uniquely barcoded molecules relative to the number of reads produced, read families will tend to be too small for the formation of consensus sequences. Alternately, if there are too few uniquely barcoded molecules relative to the reads produced, read families will be much larger than they need to be, wasting reads. To explore this dependence further, we applied duplex barcodes to sheared yeast genomic DNA and used five different concentrations of input molecules for amplification. Each sample was amplified under identical conditions and sequenced, with the same number of total sequencing reads requested for each. We

measured the efficiency with which standard barcoding and duplex barcoding consensus sequences were formed across the five datasets (Figure 2-5B). For both standard and duplex barcoding, the efficiency rose, peaked, and declined within the range of library sizes used. The efficiency peaked at a very small library size of 4 attomol for both standard barcoding and duplex barcoding (7.8% and 1.3%, respectively). Although concentration of input DNA is easy to control in setting up these reactions, the optimal library size depends on both barcoding efficiency and the number of reads actually produced in the final dataset. Therefore, the initial input library size must be empirically determined for each experiment.

A.

	Bases In	Bases Out	Efficiency
Circle Seq	413M	83M	20.2%
Barcoding^S	2,210M	66M	3.0%
Duplex Bar^S	2,210M	18M	0.8%

B.

Input DNA Molecules	Eligible Reads	Barcoding		Duplex Barcoding	
		# Read Families	Efficiency	# Read Families	Efficiency
4000 amol	304k	11	0.004%	0	0.0%
400 amol	351k	978	0.3%	0	0.0%
40 amol	254k	18,020	7.1%	236	0.1%
4 amol	372k	29,036	7.8%	4,651	1.3%
0.4 amol	431k	2,906	0.7%	495	0.1%

Figure 2-5: Efficiency of circle sequencing and barcoding methods. (**A**) The table shows key metrics of efficiency for the three approaches discussed: circle sequencing, standard barcoding, and duplex barcoding. "Bases in" refers to the total number of bases used to build read families. For the barcoding-based approaches, these are bases in well-formed, uniquely mapping reads. For circle sequencing, these are bases in reads showing clear periodicity. "Bases out" refers to consensus bases. Consensus bases are produced from read families with at least three members (at least three members derived from each strand for duplex barcoding). Efficiency is calculated as the number of consensus bases produced divided by the total number of bases used to produce them ("bases out" divided by "bases in"). Standard and duplex barcoding values (S superscript) are derived from a dataset from ref. (7), which was reanalyzed here. (**B**) Standard barcoding and duplex barcoding were used to sequence yeast genomic DNA. Tenfold serial dilutions of the input material were made before the library amplification step and an 18-cycle PCR was performed. The number of eligible reads refers to the number of reads used to build read families. Also shown are the number of read families consisting of at least three members (standard barcoding) or at least three members from each strand (duplex barcoding), and the efficiency of consensus sequence formation (ratio of read families produced to total eligible reads).

Next, we looked at the distribution of sizes of read families for the two dilutions producing the highest efficiency (40 and 4 attomol) (Figure 2-5B). These datasets produced 18,020 and 29,036 read families with 3 or more members (Figure 2-5B). In the 40 attomol library, most read families contained only one or two members (Figure 2-6). In the 4 attomol library, most read families have many more than 3 members, with the average read family size being ~12 (Figure 2-6). These results clearly demonstrate a direct correlation between input concentration and efficient use of sequencing reads to produce consensus sequences.

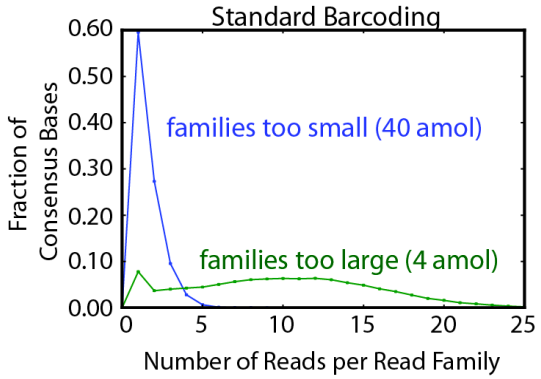


Figure 2-6: Read family sizes in the standard barcoding method. The distribution of sizes of read families (number of reads per read family) produced by standard barcoding with 40 attomol input (blue) and 4 attomol input (green).

To determine the precise expected relationship between input library size and efficiency beyond the five experimental points sampled, we analyzed an idealized theoretical model of the barcoding process. In this model, every barcoded input molecule is massively and uniformly amplified so that each sequencing read produced has an equal and independent chance to sample each of the original input molecules. For simplicity, we assume that exactly 1 million usable sequencing reads are always produced. The expected efficiencies of recovering input molecules at least three times for standard

barcoding (Figure 2-7, purple) and recovering both strands of input molecules at least three times each for duplex barcoding (Figure 2-7, green) are plotted as a function of the number of uniquely barcoded input molecules. As discussed above, the idealized perfect efficiency for these approaches would be 33% (or half of this for duplex barcoding). However, unavoidable variance in the distribution of read-family sizes due to the random sampling process caps the efficiency of standard barcoding at 19% and duplex barcoding at 8%. Perhaps more notable is the rapid decline in efficiency observed when the number of barcoded input molecules falls outside a narrow range around these peaks. In practice, precise control over the ratio of barcoded molecules to usable sequencing reads can be difficult to achieve. For instance, inferred estimates of the fraction of input molecules that had barcodes successfully ligated to them varied by a factor of 1.7 across the experiments that we analyzed (Table 2-1). There is also substantial run-to-run variability in sequencing machine output and in the number of reads wasted on undesired products such as adapter dimers or ill-formed barcodes. In summary, barcoding-based efficiencies are difficult to control and capped at an absolute upper bound of 19%.

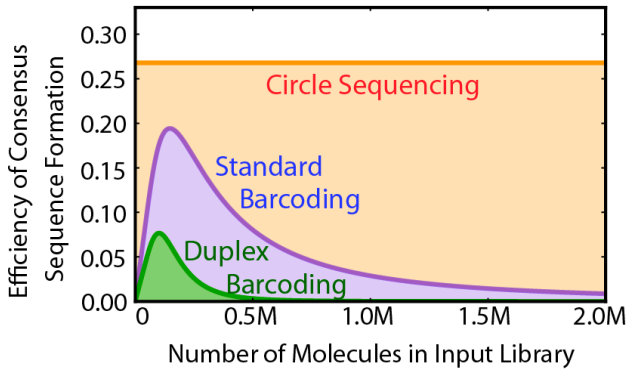


Figure 2-7: Predicted circle sequencing efficiency is independent of library size. Theoretical efficiency of consensus sequence formation from 1,000,000 sequencing reads using standard barcoding (purple), duplex barcoding (green), and circle sequencing (orange) as a function of the number of unique molecules in the input library.

	Input DNA Molecules[a]	Inferred Molecules[b]	Barcoding Efficiency[c]
Standard Barcoding	2,408,856,600	12,055,808	0.50%
	240,885,660	1,404,216	0.58%
	24,088,566	149,068	0.62%
	2,408,856	13,750	0.57%
	240,885	1,480	0.61%
Standard[S]	24,088,566	210,789	0.88%

[a] number of total molecules (barcoded and unbarcoded) in the PCR reaction.
[b] number of barcoded molecules calculated using the total number of reads and number of unique sequences acquired.
[c] efficiency of barcoding = inferred molecules/input DNA molecules.
[S] data from Schmitt *et al.*

Table 2-1: Efficiencies of barcode ligation.

We next examined the theoretical efficiency of circle sequencing. The expected efficiency for 150-base circular templates sequenced using 2 x 250 base reads is 27% (Figure 2-7). This number is less than the predicted efficiency of 33% because paired end reads are used. Because rolling circle amplification products are sheared randomly and read from either end, each read from a read pair will begin at a different base position within the repeated sequence. This offset introduces some variability in the number of repeats in a read family, with not all repeats being full length (illustrated in Figure 2-2, step 4). Importantly, however, because the repeats within a read family are physically linked and do not need to be recovered from a bulk mixture by sampling, efficiency will not vary with the number of molecules in the input library. Efficiency is therefore a flat line across all input library sizes. This plot demonstrates two key features of circle sequencing: the theoretical peak efficiency is higher than for barcoding-based approaches, and this efficiency is insensitive to experimental conditions, sidestepping a major liability of barcoding-based approaches.

	Input Concentration	Total Reads	Yield	Error Rate
Standard Barcoding	4000 amol	567,396	0.00%*	ND
	400 amol	631,578	0.1%	7.04×10^{-6}
	40 amol	476,885	3.3%	1.19×10^{-5}
	4 amol	621,151	3.4%	1.09×10^{-5}
	0.4 amol	814,987	0.4%	3.88×10^{-4}
Circle Sequencing	493 fmol	1,672,718	8.2%	9.17×10^{-6}
	493 fmol	1,074,244	9.9%	7.79×10^{-6}
	154 fmol	597,419	8.5%	9.30×10^{-6}
	15.4 fmol	717,337	8.0%	1.04×10^{-5}
	1.54 fmol	658,765	8.2%	1.02×10^{-5}
Standard[S]	40 amol	77,834,681	2.6%	4.61×10^{-5}
Duplex[S]	40 amol	77,834,681	0.07%	1.72×10^{-6}

* Error rate could not be calculated due to poor yield (yield of 9×10^{-6} indicated by asterisk, error rate indicated as ND).
[S] data from Schmitt *et al.*

Table 2-2: Yield and error rates of libraries analyzed.

Although the efficiency of consensus sequence formation is critically important, a wide range of other practical issues also affects the total amount of usable data produced. We define yield as the total number of high-quality consensus bases produced divided by the raw number of sequenced bases before any filtering or data processing is performed. This metric considers the overall loss of data in a sequencing project from start to finish, including not only loss due to consensus formation, but also losses due to data filtering and trimming schemes and reads that can't be mapped uniquely to the genome being sequenced. Based on this final point, the parameter of yield will therefore be somewhat genome-specific, as repetitive information and missing regions in genome assemblies can vary from genome to genome. To quantify the cumulative impact of all of these effects on the yield of error correcting methods, we used circle sequencing to sequence a set of yeast genomic libraries that varied in input concentration and/or total reads produced, and compared the data with the set of yeast genomic libraries sequenced with barcoding. The

input molecules used and the total sequencing reads obtained for each sample are summarized in Table 2-2.

Figure 2-8 shows four standard barcoding samples and five circle-sequencing samples on a plot of yield versus error rate. All five of the circle-sequencing samples, regardless of molecules in the original library or reads produced, had a yield and an error rate that clustered within a tight range (orange points). The barcoding libraries were more disperse on this plot, with the samples varying significantly in both yield and error rate (green points). Even the most efficient barcoding samples (4 attomol and 40 attomol libraries) had a yield that was only 1/3 that of the circle-sequencing samples, equating to a cost that would be three times as high for the same amount of high-quality, error-corrected data.

Although yields obtained in experiments targeting genomes of different complexities are not directly comparable for the reasons discussed above, we also include values for the barcoded M13mp2 phage genome dataset, which was produced with the Illumina HiSeq machine (7). We find that the metrics of yield and error rate are similar despite differences in the genomes sequenced and sequencing platform used (Figure 2-8, gray points). The method that currently produces the lowest error rate is the duplex barcoding method of Schmitt *et al*. (7). However, the yield of this method is very low, with ~1 out of 1,000 bases sequenced being recovered as a consensus base. For all sequencing projects, high yield and low error rate are desirable, and so the best methods will fall in the upper right hand corner of the plot in Figure 2-8 (purple arrows). Circle sequencing produces high yield and low error rate and is highly robust to experimental design in ways that barcoding approaches are not.

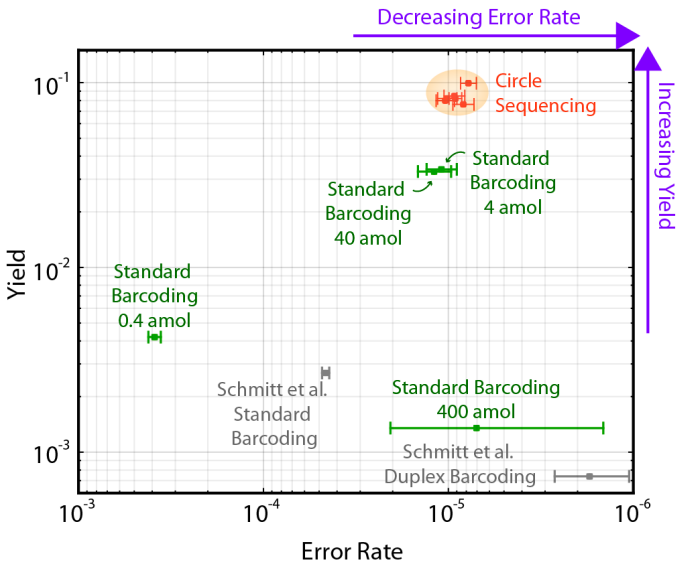


Figure 2-8: Comparison of overall yield and error rate for all error-correction methods. The yeast genome was sequenced with standard barcoding (green points) or circle sequencing (orange points) while varying input DNA concentration and/or reads produced. Error rate (x axis) is defined as the fraction of consensus bases that differ from the reference sequence. Yield (y axis) is the total number of consensus bases produced divided by the raw number of bases sequenced. Circle sequencing produces consistent error rates and yields across a range of experimental conditions (orange shading). Standard barcoding produces highly variable error rates and yields. Another library discussed in the text, from the M13mp2 phage genome generated in ref. (7), was also analyzed (gray points).

Finally, we considered whether sequence-specific biases affect our library preparation, such as bias in template circularization. This bias could result in non-uniform coverage of the genome being sequenced. As might be expected, we did observe that some degree of template bias is introduced by both barcoding and circle sequencing when each is compared with standard Illumina sequencing. This effect was slightly larger for circle sequencing although the skews in coverage were not extreme in either case (Figure 2-9). We also considered that some circular templates might have sequence features that lead to biased amplification during library preparation. This bias could result in many reads deriving from the same circular template. However, we found that greater than

98.8% of the consensus sequences produced in each dataset were derived from unique circular templates, and no single circular template produced more than five consensus sequences.

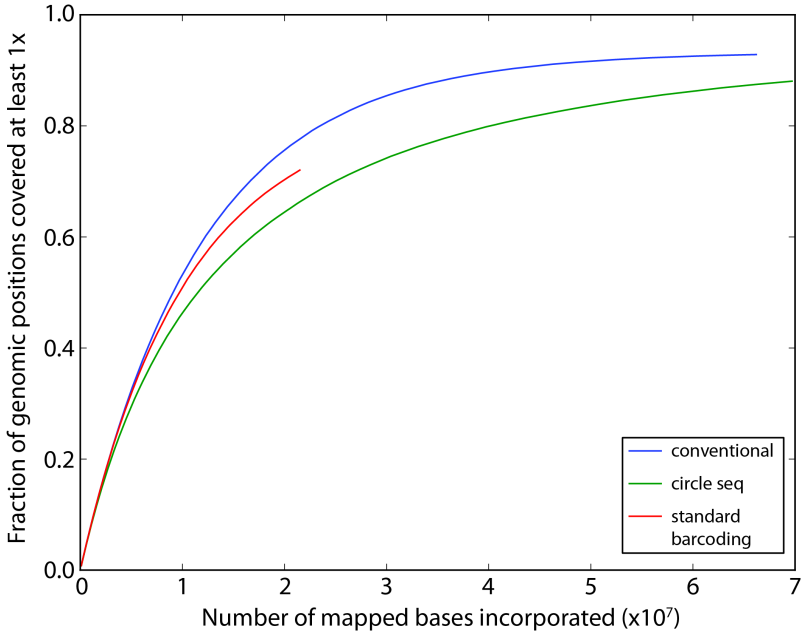


Figure 2-9: Genome coverage obtained with three different sequencing methods. A comparison of genome coverage is made between conventional MiSeq data (blue line), circle-sequencing data (green line), and standard barcoding data (red line). Here, we analyzed the observed fraction of all positions in the yeast genome (y axis) as a function of the number of processed bases used (x axis). Uniquely mapping reads (conventional) or uniquely mapping consensus sequences (circle sequencing and standard barcoding) were shuffled and then processed one-by-one while recording the fraction of all positions in the yeast genome covered at least once. To control for the influence of sequence lengths on mappability, conventional reads and barcoding consensus sequences were trimmed to exactly match the empirical distribution of circular consensus sequence lengths. The gap between the blue and green lines indicates that the enzymatic steps involved in circle sequencing introduce some degree of bias in which regions of the genome tend to be observed, but that the effect is not extreme. The difference in fraction of genomic positions covered (y axis) between the blue and green lines is never more than 11.5% (meaning that. for a given number of sequencing bases used, conventional sequencing has covered at most 11.5% more genomic positions than has circle sequencing). The barcoding line (red) terminates early because the barcoding process produced substantially less consensus sequence data.

DISCUSSION

Circle sequencing is a library preparation method for high-throughput sequencing that achieves low error rate and high efficiency. Its biggest strength is that it is efficient over a range of experimental designs (input library types and reads produced). The choice of library preparation method will ultimately be dependent on the task at hand. Standard high-throughput sequencing, combined with sufficient read depth, may still be the best choice for genome-sequencing projects. One barcoding approach, duplex barcoding, has an error rate lower than any other method because the inclusion of information from both strands of each DNA duplex helps to eliminate the effects of both jackpot mutations and damaged bases (7). Although this method is highly inefficient, duplex barcoding may be the method of choice in cases where single mutations, such as individual damaged bases in a population of DNA, must be detected (i.e., projects involving the rarest of rare variants). However, for many rare variant problems, circle sequencing would be a better choice than barcoding-based methods. Circle sequencing should be especially powerful in applications related to cancer profiling, immuno- genetics, microbial diversity, and environmental sampling.

Superficially, our method appears to resemble the SMRTbell approach of Pacific Biosciences (29). This single-molecule technology also circularizes DNA and uses redundant information produced as a polymerase repeatedly traverses a circular template to reduce error. A key conceptual difference is that our method produces intermediate physical products containing multiple copies of the information in the templates. These products can then be sequenced on platforms that offer dramatically higher throughput and per-base-call accuracy than the single-molecule platform of Pacific Biosciences. We have successfully implemented our method using 2 x 150 base and 2 x 250 base reads on the Illumina MiSeq machine; in principle, appropriately sized circles could be used with

2 x 100 base reads currently available on the higher-throughput HiSeq machine. One technical point to consider in the application of our method is that circle sequencing products, because of the repetitive nature of the information contained, might be especially prone to problems in the clonal amplification that takes place on some high-throughput sequencing machines. Although we did detect this phenomenon, we estimate that the effect is small.

CONCLUSIONS

We show that the circle sequencing method described in this chapter has an rate of 7.6 x 10^{-6} per base sequenced, dramatically improving the error rate of Illumina sequencing and putting error on par with low-throughput, but highly accurate, Sanger sequencing. In addition, circle sequencing also had substantially higher efficiency and lower cost than existing barcode-based schemes for correcting sequencing errors. Currently, we are in the process of improving this method by constructing circularized templates that link together both strands of double-stranded input molecules. By incorporating the key insight of Schmitt *et al.*'s duplex barcoding method (7), this modification could protect against errors caused by damaged bases in starting templates while retaining the efficiency advantages of circle sequencing.

Chapter 3: Rapid evolution of *BRCA1* and *BRCA2* in primates

INTRODUCTION

Defects in the *BRCA1* or *BRCA2* genes are responsible for most hereditary forms of breast cancer and account for as many as 10% of all breast cancer cases (30). Women with a strong family history of cancer who possess a harmful *BRCA1* or *BRCA2* allele are at high risk for developing breast cancer within their lifetime (80% and 60%, respectively) (31, 32). In addition, *BRCA1* mutation carriers have a 30-40% chance of developing ovarian cancer, while *BRCA2* mutations also increase the risk of ovarian, pancreatic, prostate, and male breast cancer (31). Cancers occur when heterozygous individuals experience a somatic loss of heterozygosity event at the *BRCA1* or *BRCA2* locus, leaving only the abnormal allele intact. Because both gene products play a critical role in key cellular processes such as DNA repair, cell cycle control, and transcriptional regulation, it is clear why inactivating mutations are so detrimental. The importance of these proteins is further evidenced by the fact that both *BRCA1* and *BRCA2* null mice are embryonic lethal (33).

Given their indispensible functions in maintaining the integrity of the genome, one might expect strict evolutionary conservation of *BRCA1* and *BRCA2* over time. Indeed, some regions of *BRCA1* have experienced purifying selection strong enough to operate even on synonymous mutations (34). However, contrary to this line of reasoning, a number of groups have documented the rapid evolution of *BRCA1* (12-17) and *BRCA2* (12) in mammals. Rapid evolution occurs when a gene experiences positive natural selection for new, advantageous mutations that arise in a population. Because advantageous mutations commonly involve a change in protein sequence (non-

synonymous mutations), recurrent rounds of positive selection in a gene lead to rapid evolution of the encoded protein sequence over time. For *BRCA1*, the evolutionary rate was particularly elevated on the branches leading to humans and chimpanzees (*Pan troglodytes*) (13). The identification of this signature in *BRCA1* suggests that some alleles and polymorphisms currently circulating within the human population may offer a selectable advantage. However, both the cause and consequence of this unexpected mode of evolution seen in *BRCA1* remain unknown.

Here, we report an extensive evolutionary analysis of the primate *BRCA1* gene. In previous studies of *BRCA1* evolution, only exon 11 was examined with a limited number of primate species included in the analyses (12-17). To extend previous studies, we have generated full-length *BRCA1* sequences for 17 additional primate species. Using this more extensive dataset, we validate the finding of positive selection in humans and their closest ape relatives (in our study, chimpanzees and also bonobos (*Pan paniscus*)). We also show that specific codons in *BRCA1* have experienced recurrent positive selection over evolutionary time, both within and outside of exon 11, resulting in a small number of highly variable residue positions in an otherwise highly conserved protein. In addition, we sequenced exon 11 of *BRCA1* from populations of chimpanzee, bonobo, and rhesus macaque (*Macaca mulatta*) individuals and found that several unique polymorphisms exist within these populations. Two polymorphisms in the chimpanzee population were found to be in Hardy-Weinberg disequilibrium suggesting that selection may still be operating on this gene in modern times. Lastly, exon 11 of *BRCA2*, another important genetic determinant for hereditary breast and ovarian cancers, was also sequenced from diverse primate species. This gene also bears the surprising signature of positive selection. It is unclear why these critical genes bear this unusual evolutionary signature,

but we present one possible hypothesis involving interactions between DNA repair proteins and viruses.

MATERIALS AND METHODS

Non-human primate samples

Of the 44 chimpanzee samples evaluated in this study, 34 were obtained from the Chimpanzee Biomedical Research Resource (NIH8U42OD011197-13), which is supported through a cooperative agreement with the National Institutes of Health (NIH). This NIH-supported colony is housed at the MD Anderson Cancer Center's Michale E. Keeling Center for Comparative Medicine and Research (KCCMR) in Bastrop, TX. The origins of the chimpanzees comprising the KCCMR colony are highly diverse with only a few closely related (siblings/offspring) animals in the colony. Blood from 34 chimpanzees was collected directly into PAXgene Blood RNA Tubes (PreAnalytix) at the same time other blood samples were obtained as part of the prescheduled annual veterinary exam for each animal. Another 10 chimpanzee genomic DNA samples were purchased from Coriell.

All 44 rhesus macaque samples evaluated in this study were obtained from animals housed at the KCCMR in collaboration with researchers at this institution. The colony at the KCCMR is a closed breeding colony comprised of approximately 980 rhesus macaques of Indian-origin that originated from a colony of 286 founder animals in 1988. Blood from these animals was collected directly into PAXgene Blood RNA Tubes (PreAnalytix) at the same time other blood samples were obtained as part of the prescheduled annual veterinary exam for each animal.

Bonobo genomic DNA samples were obtained from the integrated primate biomaterials and information resource (IPBIR) of the Coriell Institute or extracted from blood samples obtained from the Columbus zoo and the Language Research Center, Georgia State University. All seven individuals are unrelated.

The remaining non-human primate samples were acquired as cell lines purchased from the Coriell Institute under a U.S. Fish and Wildlife Service permit. This study was approved by the University of Texas at Austin Institutional Review Board.

Primate *BRCA1* and *BRCA2* sequencing

Human BRCA1 and BRCA2 coding sequences were obtained from GenBank (accession number NM 007294 and NM 000059, respectively). *BRCA1* and *BRCA2* sequences from chimpanzee, gorilla, orangutan, rhesus macaque, and marmoset were obtained using the BLAT alignment tool on the UCSC genome database (http://genome.ucsc.edu/). For the remaining 18 primate sequences, primary or immortalized cell lines were grown in standard media supplemented with 15% fetal bovine serum at 37°C and 5% CO2. Cells were collected and RNA was extracted using the AllPrep DNA/RNA kit (QIAGEN). cDNA libraries were generated using SuperScript III First-Strand Synthesis Kit (Invitrogen) using oligo dT or random hexamer primers. PCR products were generated using PCR SuperMix High Fidelity (Invitrogen) and directly sequenced or cloned into pCR4 for sequencing. These sequences have been deposited in GenBank (accession numbers KM017616-KM017652).

Blood from rhesus macaque and chimpanzee individuals was collected in PAXgene Blood RNA Tubes (PreAnalytiX). RNA was extracted using the PAXgene Blood miRNA Kit (QIAGEN) and genomic DNA was obtained using the AllPrep

DNA/RNA kit (QIAGEN). BRCA1 Exon 11 was amplified from extracted genomic DNA (chimpanzee, bonobo, and rhesus macaque) using PCR SuperMix High Fidelity (Invitrogen) and sequenced.

PAML analysis

A multiple sequence alignment was generated for *BRCA1* and *BRCA2* using ClustalX2.1 (35). The alignments are straight-forward with only a few small indels. Gene sequences at each ancestral node were reconstructed using the codeml program in PAML 4.3 (36). dN/dS values along each branch of the phylogenetic tree were calculated using the free-ratio model. Substitution counts given along specified branches are the estimates made in the free ratio model, but were also calculated by directly comparing the predicted ancestral and the known extant sequences and counting differences manually. Both methods yielded the same values. The one-ratio and two-ratio models were performed as described previously (37). To detect selection, multiple alignments were fit to the NSsites models M1a (null model, codon values of dN/dS are fit into two site classes, one with value between 0 and 1, and one fixed at dN/dS = 1), M2a (positive selection model, similar to M1a but with an extra codon class of dN/dS > 1), M7 (null model, codon values of dN/dS fit to a beta distribution bounded between 0 and 1), M8a (null model, similar to M7 except with an extra fixed codon class at dN/dS = 1), and M8 (positive selection model, similar to M7 but with an extra class of dN/dS > 1). Model fitting was performed with multiple seed values for dN/dS (ω) and assuming either the f61 or f3x4 model of codon frequencies (38). Likelihood ratio tests were performed to assess whether permitting some codons to evolve under positive selection gives a significantly better fit to the data than models where positive selection is not allowed (39, 40). These different

model comparisons represent different trade-offs between power and accuracy (41). In all cases the positive selection model was a significantly better fit ($p < 0.05$), and individual codons assigned to the $dN/dS > 1$ class with high posterior probabilities ($P > 0.85$ by Bayes Emperical Bayes (42)) were analyzed. The crystal structure was obtained from the RCSB Protein Data Bank (http://www.pdb.org) and residues under positive selection were mapped using MacPyMol (http://www.pymol.org).

Hardy-Weinberg equilibrium test

Single nucleotide polymorphisms (SNPs) were annotated for each bonobo, chimpanzee, and rhesus macaque individual. Allele frequencies were calculated for each SNP and tested for departure from Hardy-Weinberg equilibrium (http://www.oege.org) (43). Chi squared values were calculated using 1 degree of freedom. A p-value (after Bonferroni correction) < 0.0056, 0.0056, and 0.0042 for bonobos, chimpanzees, and rhesus macaque, respectively, was considered statistically significant.

RESULTS

BRCA1 is evolving under positive selection in primates

To expand our understanding of the positive selection shaping *BRCA1* in primates, we obtained cell lines from 17 simian primate species, harvested total RNA, and created cDNA libraries. From these, the 5.6 kilobase full-length coding region of *BRCA1* was sequenced. These sequences were combined with full-length *BRCA1* sequences from six primate species with available genome projects, creating an alignment

of 23 full-length *BRCA1* sequences. 17 out of the 23 full-length sequences have never before been analyzed (asterisks in Figure 3-1).

The type of selection that a gene has experienced can be inferred from its rate of accumulation of non-synonymous (changing the encoded amino acid; denoted dN) and synonymous (silent; dS) substitutions over time. Protein-altering mutations are far less likely to be tolerated than synonymous mutations, and so $dN/dS << 1$ for the vast majority of genes encoded by human and other mammalian genomes (44). Some genes, such as pseudogenes, evolve neutrally with $dN/dS \sim 1$ because there is not strong selection for or against new mutations in these genes. Finally, selection in favor of non-synonymous mutations results in a $dN/dS > 1$. These genes are classified as being under positive selection, and are experiencing continued selection for "innovation" at the protein sequence level. In these genes, not only has the penalty against protein-altering mutations been relaxed, but this very type of mutation is being selectively retained. Using PAML (45), we fit the full-length *BRCA1* alignment to models of positive selection where a subset of codons is allowed to evolve with $dN/dS > 1$ (M2a, M8) and to null models not allowing positive selection (M1a, M7, M8a). Likelihood ratio tests revealed that the dataset fit the positive selection models significantly better than the null models ($p < 0.05$, Table 3-1). Thus, *BRCA1* has experienced selection in favor of non-synonymous mutations over the speciation of simian primates.

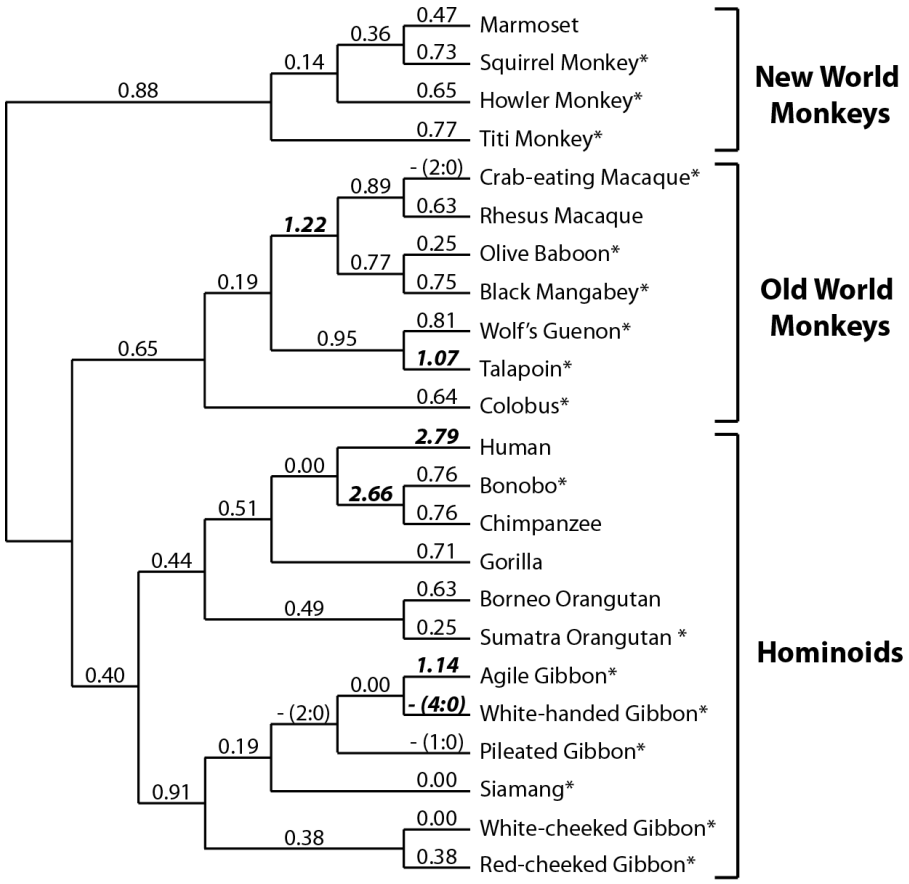


Figure 3-1: Evolution of *BRCA1* over the course of primate speciation. dN/dS values for each branch of the primate phylogeny were calculated using the free-ratio model in PAML (45). Branches exhibiting dN/dS values > 1 are shown in bold italics. Dashes (-) represent branches where zero synonymous substitutions are predicted to have occurred. On these branches, dS = 0 and dN/dS can therefore not be calculated. In these instances, the numbers of non-synonymous (N) and synonymous (S) substitutions predicted to have occurred along each branch are indicated in parentheses (N:S). Of these, branches that experienced 4 or more non-synonymous changes are in bold italics. Asterisks indicate new sequences generated in this study.

M1a-M2a	ω_0[a]	codon freq.[b]	2ΔlnL[c]	df[c]	p-value[c]	M1a-M2a	ω_0[a]	codon freq.[b]	2ΔlnL[c]	df[c]	p-value[c]
BRCA1	0.4	f61	10.0	2	0.0066	***BRCA2***	0.4	f61	21.3	2	<0.0001
	0.4	f3x4	6.1	2	0.0466		0.4	f3x4	16.1	2	0.0003
	1.6	f61	10.0	2	0.0066		1.6	f61	21.3	2	<0.0001
	1.6	f3x4	6.1	2	0.0466		1.6	f3x4	16.1	2	0.0003
M7-M8	ω_0[a]	codon freq.[b]	2ΔlnL[c]	df[c]	p-value[c]	**M7-M8**	ω_0[a]	codon freq.[b]	2ΔlnL[c]	df[c]	p-value[c]
BRCA1	0.4	f61	10.6	2	0.0049	***BRCA2***	0.4	f61	23.3	2	<0.0001
	0.4	f3x4	6.2	2	0.0447		0.4	f3x4	18.6	2	<0.0001
	1.6	f61	10.6	2	0.0049		1.6	f61	23.3	2	<0.0001
	1.6	f3x4	6.2	2	0.0447		1.6	f3x4	18.6	2	<0.0001
M8a-M8	ω_0[a]	codon freq.[b]	2ΔlnL[c]	df[c]	p-value[c]	**M8a-M8**	ω_0[a]	codon freq.[b]	2ΔlnL[c]	df[c]	p-value[c]
BRCA1	0.4	f61	10.1	1	0.0015	***BRCA2***	0.4	f61	19.9	1	<0.0001
	0.4	f3x4	6.2	1	0.013		0.4	f3x4	15.1	1	0.0001
	1.6	f61	10.1	1	0.0015		1.6	f61	19.9	1	<0.0001
	1.6	f3x4	6.2	1	0.013		1.6	f3x4	15.1	1	0.0001

[a] Initial seed value for ω (dN/dS).

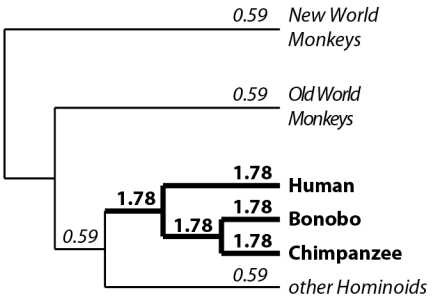
[b] Model of codon frequency.

[c] Twice the difference in the natural logs of the likelihoods (2 × ΔlnL) of the two models being compared (a model that allows positive selection (M2a or M8) is compared to a null model (M1a, M7, M8a)). This value is used in a likelihood ratio test along with the degrees of freedom (df). The p-value indicates the confidence with which the null model can be rejected.

Table 3-1: PAML analysis of *BRCA1* and *BRCA2*.

We next estimated dN/dS values on each branch on the primate evolutionary tree using the free-ratio model in PAML. As expected, most branches exhibited a dN/dS < 1 (Figure 3-1). The branch leading to humans had the most elevated signal with a dN/dS of 2.79. The second highest value of dN/dS on the *BRCA1* tree is found on the branch leading to the last common ancestor of bonobos and chimpanzees, with a dN/dS of 2.66. Because the free-ratio model is highly parameterized, we next compared one-ratio and two-ratio models to determine whether selection has differentially affected the human, chimpanzee, and bonobo clade. As shown in Figure 3-2, our simian primate dataset fit the two-ratio model significantly better than the one-ratio model, with the human, chimpanzee, and bonobo clade exhibiting a dN/dS of 1.78, while all other branches had a dN/dS of 0.59. In summary, our extended primate dataset shows that *BRCA1* is

experiencing positive selection, and that the most intense selection has operated on the human/chimpanzee/bonobo clade.



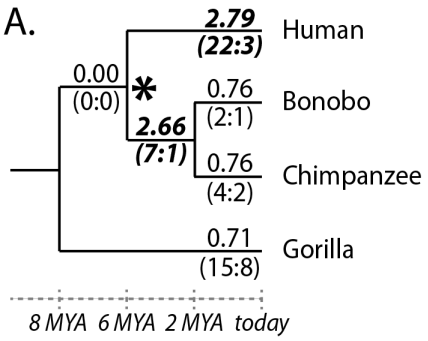
Model	2ΔlnL	df	*p-value*	ω_0	ω_1
one-ratio ($\omega_0=\omega_1$)	*8.5*	*1*	*0.0036*	*0.62*	ω_0
two-ratio ($\omega_0\neq\omega_1$)				*0.59*	*1.78*

Figure 3-2: *BRCA1* has been evolving differentially during primate speciation. The human, bonobo, and chimpanzee clade was isolated and dN/dS values were calculated using the one-ratio and two-ratio models in PAML. The two-ratio model was a better fit as determined by the likelihood ratio test shown in the box. ω0 is the calculated dN/dS for all branches under the one-ratio model, or for background branches under the two-ratio model, and ω1 is the dN/dS for the isolated branches in the two-ratio model.

Based on a comparison of extant and predicted ancestral sequences, humans are estimated to have accumulated 25 substitutions in the *BRCA1* gene since their divergence from chimpanzees and bonobos six million years ago, 22 of which are non-synonymous (Figure 3-3A). In order to understand how unusual this is, we looked at the evolution of other genes, specifically ones encoding BRCA1-interacting proteins, along the branch leading to humans. Because we do not have extended sequence sets for all of these genes, we took a simpler approach. For each gene, we aligned the human, chimpanzee, and gorilla sequences and manually counted the number of human-specific substitutions (any position where the human gene sequence differs from both the chimpanzee and gorilla gene sequence). These were categorized as non-synonymous (N) or synonymous (S)

based on how they affected the codon in which they were found. When these values are normalized to gene size, *BRCA1* has the highest enrichment of non-synonymous substitutions [(N/kb) / (S/kb)]. Care must be taken in comparing this metric between genes, because different genes have different equilibrium codon frequencies, and therefore have different mutational opportunities for synonymous and non-synonymous mutations. However, the *BRCA1* gene has an enrichment ratio that is more than 4-fold higher than any of the other genes shown (Figure 3-3B).

A.



B.

Gene	kb	N	S	N/kb	S/kb	N/kb / S/kb
BRCA1	*5.589*	*22*	*3*	*3.9*	*0.5*	*7.8*
BACH1	*2.208*	*0*	*3*	*0.0*	*1.4*	*0.0*
BARD1	*2.331*	*1*	*1*	*0.4*	*0.4*	*1.0*
CtIP	*2.691*	*5*	*7*	*1.9*	*2.6*	*0.7*
PALB2	*3.558*	*7*	*5*	*2.0*	*1.4*	*1.4*
RAD50	*3.936*	*1*	*3*	*0.3*	*0.8*	*0.4*
BRCA2	*10.254*	*7*	*8*	*0.7*	*0.8*	*0.9*

Figure 3-3: *BRCA1* evolution in the human, bonobo, and chimpanzee clade. (**A**) dN/dS values for *BRCA1* were calculated on each branch of the using the free-ratio model in PAML. dN/dS values > 1 are shown in bold italics. The numbers of non-synonymous (N) and synonymous (S) substitutions predicted to have occurred along each branch are indicated in parentheses (N:S). The asterisk represents the last common ancestor of humans, bonobos, and chimpanzees. MYA, million years ago. (**B**) The number of human-specific non-synonymous (N) and synonymous (S) substitutions in *BRCA1* and other genes encoding BRCA1-interacting proteins. The length of each gene is shown in kilobases (kb). Non-synonymous and synonymous substitutions are shown as number of substitutions per kilobases (N/kb and S/kb, respectively). An "enrichment ratio" of N/kb over S/kb was also calculated.

BRCA1 encodes a 220 kDa protein with two conserved domains: an N-terminal RING domain and two tandem C-terminal BRCT domains (Figure 3-4). The RING domain has E3 ubiquitin ligase activity that is essential in the DNA damage response. The BRCT motifs function as a protein-protein interaction module that binds

phosphorylated proteins involved in DNA repair, cell cycle control, chromatin remodeling, and transcription. There is also a coiled-coil region between these two domains. Interestingly, all but one of the non-synonymous substitutions predicted to have occurred in the human/bonobo/chimpanzee clade fall outside of these known structural motifs (Figure 3-4).

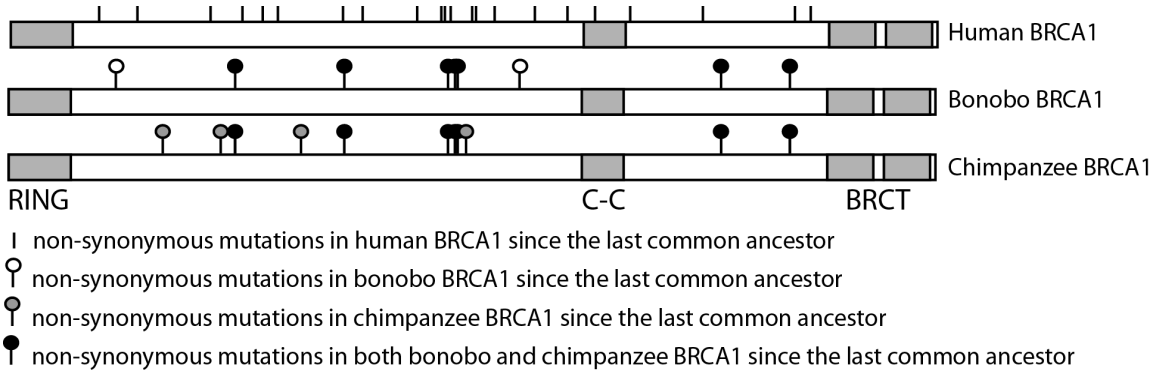


Figure 3-4: Substitutions in the *BRCA1* gene of human, bonobo, and chimpanzee. A domain diagram of BRCA1 is shown with the RING domain, coiled-coil domain (C-C), and BRCT domains indicated. On this are superimposed all of the non-synonymous substitutions predicted to have occurred in the tree shown in panel A since the divergence of humans, bonobos, and chimpanzees from their last common ancestor (asterisk in A). Vertical lines indicate substitutions specific to humans, lines with white circles are substitutions specific to bonobos, and lines with grey circles are substitutions specific to chimpanzees. Lines with black circles indicate substitutions common to both bonobos and chimpanzees.

Human variation at selected sites in *BRCA1*

The M8 model allows a class of codons to evolve under positive selection (dN/dS > 1). 10 codons were identified as belonging to this class with a high posterior probability (P = 0.85 or above). These codons do not lie in the region of *BRCA1* where it was previously reported that selection might be acting against synonymous mutations (34),

potentially given rise to a false signature of dN/dS > 1. Instead, all 10 sites show high variability between primate species at the protein level, often encoding very dissimilar amino acids (first four rows in Figure 3-5). Next, these positively selected codon positions were examined for variability within the human population. The Breast Cancer Information Core (BIC, http://research.nhgri.nih.gov/bic/) is a repository of human *BRCA1* polymorphisms. Using this database, we identified single nucleotide polymorphisms (SNPs) at amino acid sites 170, 888, 890, 1203, and 1443 (Figure 3-5). At four out of these five sites (position 888, 890, 1203, and 1443), we find that some human *BRCA1* alleles encode a unique amino acid not observed in any of our primate sequences. In addition, SNPs known to cause human disease occur in six out of 10 sites. In all cases, these disease-linked SNPs are not amino acid-altering mutations, but rather more radical frame-shifting or nonsense mutations (Figure 3-5). In particular, nonsense mutations occurring in codon 1443 are among the most common mutations documented in the BIC.

	170	439*	798	835*	888	890	1203	1370*	1443*	1510
Human BRCA1	R	A	P	H	H	G	R	S	R	M
Hominoids	R/Q	A/S	P	H	H/Q	G/R	R	S	R/Q	M/L
Old World Monkeys	R/Q	P	R/L/Q	S/R/H/G	H	R/K	R/Q	S	R	V
New World Monkeys	W/R	P/A/H	P	P	H/C	G/R	Q	F/V/S	R/Q	V/M
Human SNPs	W/Q	-	-	-	Y	R/V	G/Q	-	G/Q	-
Disease Mutations	-	-	delC	ins17	-	delG	X	delTG	X	-

Figure 3-5: Specific codons in *BRCA1* have experienced positive selection during primate speciation. Shown are the ten codons that have evolved under positive selection (dN/dS > 1) in primates with a P>0.85. Codons with a P>0.95 are indicated with asterisks. The amino acids encoded at these positions in human BRCA1 are shown, along with those found in hominoids, old world monkeys, and new world monkeys. In addition, human SNPs and disease mutations also found at these sites are listed. X refers to a single nucleotide mutation that results in a termination codon.

In Figure 3-6, all 10 sites of positive selection were mapped onto a domain diagram of BRCA1 (bottom) along with the most common human non-synonymous SNPs found in the BIC (top). As described previously for mutations accumulated in the human/chimpanzee/bonobo clade, all but one of the positively selected residues (1370S in the coiled-coil domain) lie outside of any known structural motifs. In summary, the 10 codon positions identified in this analysis are highly variable between primate species and within the human population, and are involved in the etiology of cancers associated with this gene. Disease-associated SNPs at these sites tend to be radical, protein-truncating mutations. However, a presumably distinct phenomenon appears to be driving selection in favor of non-synonymous point mutations at these positions.

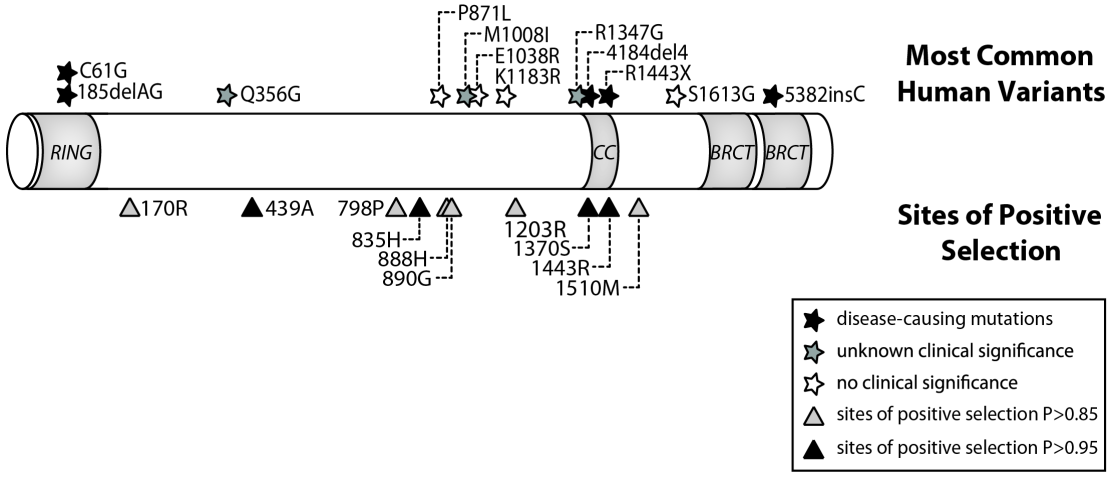


Figure 3-6: Sites of positive selection on the BRCA1 protein. A domain diagram of BRCA1 is shown with the RING domain, coiled-coil domain (CC), and BRCT domains. The triangles at the bottom represent sites of positive selection (grey - P>0.85, black - P>0.95). The 12 most common human variants recorded in the BIC are shown at the top of the diagram as stars. The black stars indicate disease-causing mutations, white stars represent variants with no known clinical significance, and grey stars are those with unknown significance.

BRCA1 variation in other primate populations

So far, we have documented sequence differences between the BRCA1 proteins of different primate species. We have shown that non-synonymous substitutions are accumulating in *BRCA1* faster than expected under constrained, or even neutral, evolution. We next wished to explore whether positive selection is still acting on *BRCA1* in modern populations. There is already evidence that this is true in the human population, because several *BRCA1* SNPs have been found to depart from Hardy-Weinberg equilibrium in European populations (46, 47) and in Australia (13). We wished to determine if the same might be true in bonobo and chimpanzee populations. We amplified and sequenced the largest *BRCA1* exon, exon 11 which is ~3.4 kilobases and comprises ~61% of the *BRCA1* coding region, from the genomic DNA of seven bonobo and 44 chimpanzee individuals (Table 3-2). In bonobos, we found nine polymorphic sites, eight of which were single nucleotide polymorphisms (SNPs), with three of these being non-synonymous. Eight of the SNPs were in Hardy Weinberg equilibrium. Interestingly, one bonobo individual was also homozygous for a seven amino acid deletion (Δ1058-1064) (Table 3-2). Hardy-Weinberg equilibrium was rejected for this polymorphism, although the support was weak and did not survive correction for multiple testing (Table 3-2). The chimpanzee sequence set revealed nine SNPs, seven of which were non-synonymous. Interestingly, in this larger sample set (n = 44), three of the non-synonymous SNPs were found to be in Hardy Weinberg disequilibrium, suggesting that selection is acting either for (E309K and G590S) or against (G1077R) these mutations. The support for one of these (E309K) was weak and did not survive correction for multiple testing (Table 3-2). It is particularly intriguing to see that humans also share with chimpanzees this same S/G SNP at position 590. In both the bonobo and chimpanzee populations, all synonymous SNPs were in Hardy-Weinberg equilibrium.

Species	SNPs[a]	Genotype			p-value[b]	Human[c]	Human Polymorphisms	
		AA	AB	BB			BIC[d]	1000 genomes[e]
Bonobo	I493L	6	1	0	0.841	I		
n = 7	T582M	6	1	0	0.841	T		
	L833L	4	3	0	0.471	L	dupAAGTATCCAT*	
	V1047V	5	1	1	0.128	V		
	G1048G	5	1	1	0.128	G	G1048D, G1048V, G1048G	G1048D, G1048V, G1048G
	T1051I	5	1	1	0.128	T		
	Δ1058-1064[HWD]	6	0	1	0.008			
	V1061V	6	1	0	0.841	I	delA*	delA*
	G1062G	6	1	0	0.841	G		
Chimpanzee	E309K[HWD]	19	14	11	0.023	K	K309T	K309Q, K309T
n = 44	E427K	34	9	1	0.663	E		
	S578S	40	4	0	0.752	S	S578Y	S578Y
	G590S[HWD]	20	12	12	*0.004*	S	S590G	S590G
	K731E	19	16	9	0.122	K	delAGAAG*	delAGAAG*
	I925T	34	9	1	0.663	I	I925L	I925V, I925L, insT*
	S1042S	41	3	0	0.823	S		
	G1077R[HWD]	42	1	1	*1.4E-5*	G		G1077W, G1077G
	G1100E	20	16	8	0.155	G		
Rhesus	A225A	42	2	0	0.888	A		
n = 44	N375S	43	1	0	0.920	N	delA*, N376S	delA*, N376S
	R466R	42	2	0	0.888	K	K467X*	K467X*
	T487S	43	1	0	0.920	T	insA*	insA*
	N684N	29	14	1	0.647	N		
	V739M	38	6	0	0.624	V	V740L	V740L, insA*
	D773G	29	15	0	0.173	G		
	D852D	40	4	0	0.752	D	insA*	insA*
	N923H	40	4	0	0.752	N		
	K936K	40	4	0	0.752	K		
	A1167E	40	4	0	0.752	A		
	Q1203R	29	14	1	0.647	R	R1203Q, R1203G, R1203X*	R1203Q, R1203G, R1203X*

[a] Numbering refers to the amino acid position in the respective primates. In the case of rhesus macaques, amino acids 375 to 936 correspond to amino acids 376 to 937 in humans.
[b] p-values were calculated using a chi-squared test with a df = 1. A p-value cutoff (after Bonferroni correction) < 0.0056, 0.0056, and 0.0042 for bonobos, chimpanzees, and rhesus macaque, respectively, was considered statistically significant. Tests that survived this correction have the p-value listed in italics.
[c] Amino acid found in the human BRCA1 protein at each of the positions listed.
[d] Human variants found at the positions indicated in the Breast Cancer Information Core.
[e] Human variants found at the positions indicated in the 1000 Genomes database.
* Known human disease-causing variant.
HWD SNPs found to be in Hardy-Weinberg Disequilibrium.

Table 3-2: SNP Analysis of *BRCA1* in Bonobo, Chimpanzee, and Rhesus Macaque Individuals

We also sequenced exon 11 from 44 rhesus macaque individuals. Rhesus macaques are not part of the human/chimpanzee/bonobo clade and are instead distantly-related members of the Old World monkey clade (Figure 3-1). In these macaques, we

found 12 SNPs in *BRCA1*, with seven being non-synonymous (Table 3-2). This includes a SNP found at position 1203, a site of positive selection in the inter-species dataset. This codon is also the site of a known disease-linked mutation in humans; however, the cancer-linked SNP at this position introduces a stop codon. Nonetheless, all of these are in Hardy-Weinberg equilibrium.

Caution must be used when interpreting signatures of selection acting on polymorphisms in primate populations. When sampling primates, it is not possible to get completely random and non-related population sets. Deviations from Hardy-Weinberg equilibrium may occur due to factors other than selection. Reasons for falsely rejecting Hardy Weinberg equilibrium include 1) non-random mating, 2) small population sizes which magnify the effects of genetic drift, 3) introduction of new alleles, 4) population subdivision or admixture, 5) biases in sequencing errors, and 6) linkage disequilibrium with another locus under selection. Because the chimpanzee population consists of individuals from two different subspecies, admixture could plausibly lead to rejection of Hardy Weinberg equilibrium.

We also performed the McDonald-Kreitman and Tajima's D tests on our datasets (data not shown). The tests were not significant and therefore do not support selection acting on any of these polymorphisms. False conclusions in this test can again result from a population with hidden structure. In summary, while the analyses using the simian primate dataset consisting of 23 species suggest that recurrent positive selection has been acting on *BRCA1* over the course of several million years, the Hardy-Weinberg equilibrium tests performed here and by others indicate that selection is acting on modern day humans, and possibly also chimpanzees.

***BRCA2* is also evolving under positive selection in primates**

Because of the rapidly evolving nature of *BRCA1*, we also completed an evolutionary analysis of *BRCA2*, another strong determinant for hereditary breast and ovarian cancer. Although *BRCA2* has been shown to be under positive selection, only a small number of primate species was included in this study (12). We sequenced the ~5 kilobase exon 11 from 18 primate species. Exon 11 is the largest of 27 exons and encodes about 50% of the entire BRCA2 protein. The sequences, along with six additional sequences from available genome projects, were assembled into a multiple alignment. We fit the alignment to positive selection and null models as described above. The positive selection models were again a significantly better fit to the sequence set than the null models, with a p value ≤ 0.0003 (Table 3-1). In summary, *BRCA2* is under positive selection in primates as well, although this signature appears not to be concentrated on the human/chimpanzee/bonobo clade.

In contrast to BRCA1, BRCA2 is a 390 kDa nuclear protein that is exclusively involved in the homologous recombination pathway for repairing double-strand breaks. The eight BRC motifs and the extreme C terminus mediate interactions with and recruitment of Rad51, a protein that catalyzes strand invasion during homologous recombination (48-50). All eight BRC repeats are encoded within exon 11. The M8 model estimates that five codons are evolving under positive selection with posterior probability > 0.85 (Figure 3-7A). Two of these positively selected sites were found to have a human polymorphism documented in the BIC (Figure 3-7A). When all five sites of positive selection are mapped onto a domain diagram of BRCA2 (Figure 3-7B), they cluster within the first three BRC domains (1008, 1225, and 1426) and the intervening regions (1159 and 1272).

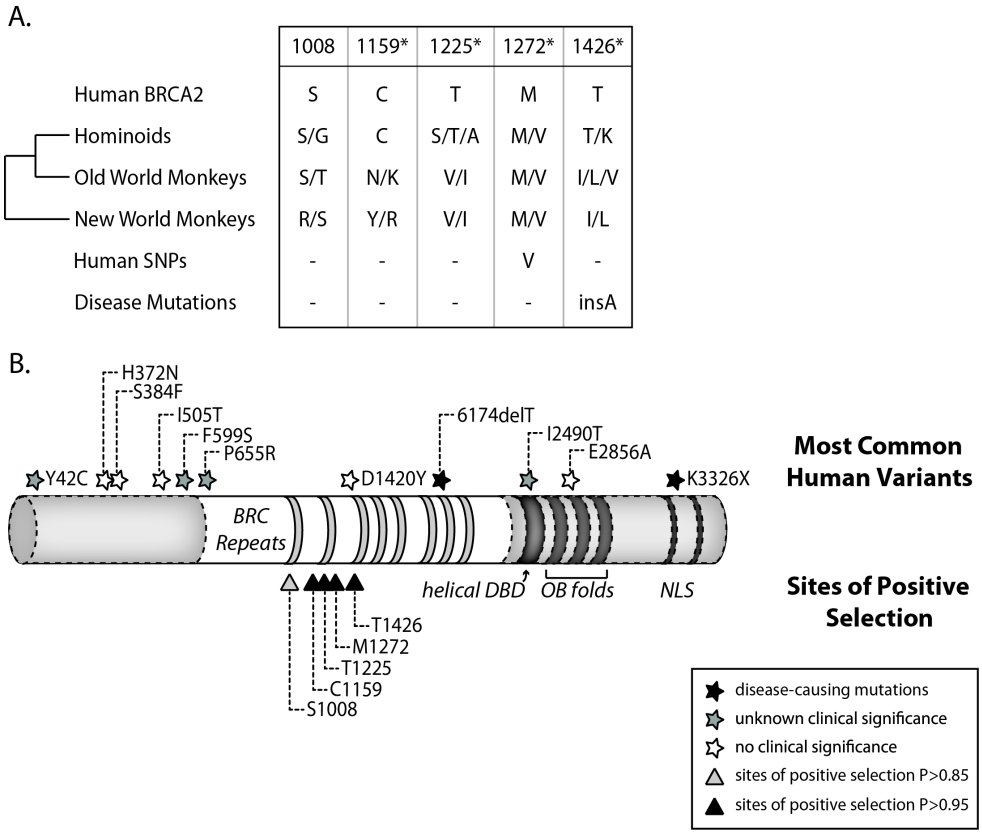


Figure 3-7: Codons in exon 11 of *BRCA2* that have experienced positive selection in primates. (**A**) 5 codons in exon 11 of *BRCA2* were found to be under positive selection in primates. All sites had a P>0.95 (indicated with asterisks) except for S1008 (P=0.9). The amino acid encoded by human BRCA2 at each of these codons is shown. The amino acids encoded by hominoids, old world monkeys, and new world monkeys are also shown. Human SNPs and disease mutations deposited to the BIC are listed at the bottom. (**B**) A domain diagram of BRCA2 is depicted with the 8 BRC repeats, helical DNA binding domain (helical DBD), OB folds, and nuclear localization signals (NLS). Only exon 11 was sequenced in this study (section in white). The sites of positive selection are represented as triangles at the bottom of the diagram. The 11 most common protein-altering variants in the BIC are marked as stars at their respective locations at the top. Black stars correspond to disease-causing mutations, white stars are variants with no known clinical significance, and grey stars are positions with unknown significance.

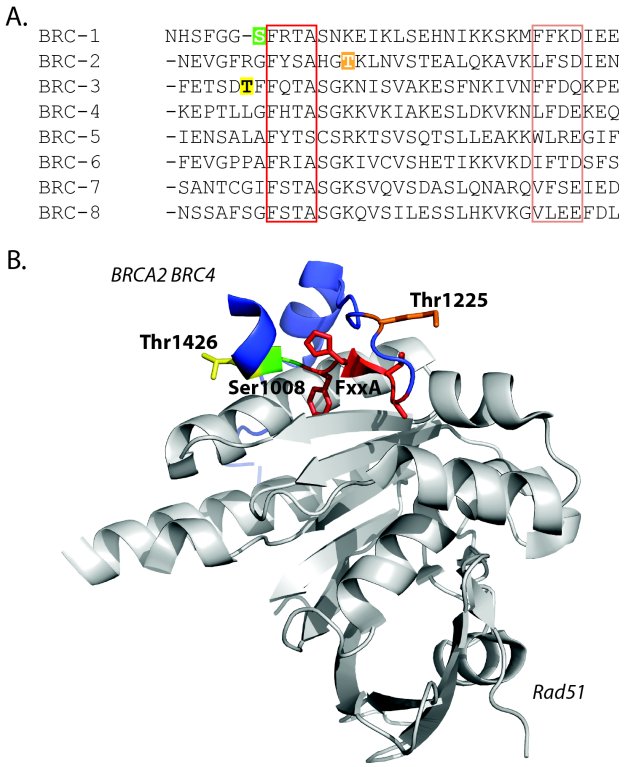


Figure 3-8: The sites of positive selection lying within the BRC repeats of BRCA2 are located adjacent to the Rad51 binding region. (**A**) The 8 BRC repeats of the human BRCA2 protein were aligned using ClustalX. The red and peach colored boxes are the motifs within the BRC repeats thought to facilitate binding with Rad51 (51). Residues 1008, 1225, and 1426 are colored in green, orange, and yellow, respectively. All three sites lie just adjacent to the FxxA motif which interacts with two hydrophobic pockets in the Rad51 oligomer. (**B**) The co-crystal structure of BRC4 (blue) in complex with Rad51 (grey) is shown (PDB ID 1N0W (52)). The FxxA motif is depicted in red. Residues 1008, 1225, and 1426 are shown in green, orange, and yellow, respectively.

To examine this further, we aligned the amino acid sequence of all eight BRC repeats of human BRCA2 and highlighted sites 1008, 1225, and 1426 (Figure 3-8A). Surprisingly, all three sites of positive selection lie adjacent to a hydrophobic motif (FxxA) known to mediate interactions with Rad51 (Figure 3-8A red box). Since the co-crystal structure of the BRCA2 BRC4 in complex with Rad51 is available, we mapped these three sites to their analogous positions in BRC4 and found that they are in close

proximity to the Rad51 binding interface (Figure 3-8B, PDB: 1N0W) (52). The clustering of these residues near this interface might provide a clue to the driver of natural selection at these sites.

DISCUSSION

Nearly all known cases of recurrent positive selection in primate genomes involve genes in one of three categories: 1) immunity, 2) environmental perception (such as odorant and taste receptors), or 3) sexual selection and mate-choice (53, 54). This is due to the fact that ever-changing external stimuli (i.e. pathogens, environmental odors/tastes, etc.) drive the selection of new allelic variants. For example, immunity factors that are constantly challenged by pathogens exhibit some of the most striking signatures of positive selection seen in primate genomes (55-60). Here, immunity genes will experience positive selection for protein-altering mutations that improve recognition of a relevant pathogen. Conversely, the pathogen will counter-evolve to escape detection, again placing selective pressure on the host population for new mutations that improve the immunity protein. This cycle can repeat itself indefinitely, resulting in an ever-escalating host-virus arms race. Therefore, it is surprising to see that *BRCA1* and *BRCA2*, genes that do not classically fit into any of the three categories listed above, are evolving in a similar manner to these highly adaptive immunity genes. In addition to the two described here, other DNA repair genes have also been shown to evolve under positive selection (11, 61), but the driver behind this unusual finding remains to be identified.

An intense battle exists between host DNA repair machinery and viruses, and we propose that this could contribute to the evolutionary signatures documented here. Many viruses are known to interact with the DNA repair machinery and cell cycle regulators

(62, 63). One fundamental issue is that the free ends of viral genomes are exposed, in contrast to the host's DNA, which is capped by telomeres. Despite this, many viruses need to access the nucleus where the host's DNA repair machinery recognizes these un-capped viral genome ends as "damaged" cellular DNA, activating the DNA damage response. In order for productive infection to proceed, viruses must actively thwart these host repair pathways. For example, DNA repair proteins interfere with the adenovirus lifecycle by concatenating the ends of newly synthesized viral DNA, inhibiting efficient packaging into viral progeny (64). In turn, adenovirus has evolved a way around this blockade by encoding proteins that mislocalize or degrade the specific host factors involved. Depending on the virus involved, host DNA repair factors can also be hijacked to facilitate viral replication. For instance, herpes simplex virus-1 simultaneously activates DNA repair constituents that aid in viral genome replication (65, 66) and counteracts those that do not (67, 68). Human immunodeficiency virus 1 is also known to activate the DNA damage response and manipulate cell cycle checkpoints through the actions of its accessory protein Vpr (69, 70). Additionally, several studies have shown that specific DNA repair proteins play critical roles in retroviral genome integration (71-74) while others seem to decrease the efficiency of infection (75-77).

One can imagine that these and other viruses that access the nucleus during replication could feasibly interact with BRCA1 or BRCA2, driving the selection of variants that ultimately lead to decreased susceptibility to infection. However, it is possible that variant alleles selected for this purpose would have detrimental consequences to protein function in the context of host DNA repair. Most of the deleterious *BRCA1* and *BRCA2* variants characterized thus far introduce stop codons or frame-shifts that result in premature truncation of the protein, the consequences of which manifest as cancer at relatively early ages. The effects of non-synonymous point

mutations, such as those documented here, might be expected to be much more subtle. The effects of subtle mutations are more difficult to assess because the resulting genomic instability may only be realized later in life and can be confounded by other genetic or environmental influences. We therefore propose a hypothesis where viruses are driving the intriguingly rapid rate of evolution seen in *BRCA1* and *BRCA2*, potentially giving rise to antagonistic pleiotropy. This would be analogous to the malaria and sickle cell anemia trade-off that is well documented (78).

CONCLUSIONS

The BRCA1 and BRCA2 proteins play key roles in the repair of damage to chromosomal DNA. We have expanded the analysis of the evolution of these genes, showing that both have been subject to recurrent positive selection during simian primate speciation. Although the force or forces driving the diversifying selection of these genes is unknown, the result is that the sequence of these proteins has been altered in humans and our closest living relatives. It remains to be seen whether this is an instance of antagonistic pleiotropy, where positive selection driven by one force causes functional consequences in another context, potentially the formation of cancers (79).

Chapter 4: A DNA repair protein constitutes a barrier to cross-species transmission of herpes simplex virus 1 in primates

INTRODUCTION

More than half of the human population and greater than 90% of adults over the age of 50 are infected with herpes simplex virus 1 (HSV-1), making it one of the most omnipresent viruses circulating in humans today (80). From an evolutionary perspective, the ability to cause mild clinical disease and establish a lifelong infection makes HSV-1 perhaps one of the most successful viruses of all time. In addition, a plethora of viral gene products is devoted to disarming or hijacking host protein functions during the course of infection. This clearly demonstrates that HSV-1 has finely tuned its interactions with its host in order to establish a delicate balance between optimal replication and survival, a product of long-term adaptation of a virus to its host (22, 44).

Many other non-human primate species also harbor their own simplex virus variants with which they have coevolved. The phylogenetic relationship between these simplex viruses mirrors that of their primate hosts, providing strong evidence for codivergence as the principal mode of primate simplex virus evolution (Figure 4-1) (81-84). Although each simplex virus species is genetically distinct from one another, enough similarities exist to allow for rare cross-species transmissions to occur in nature. One such example is the zoonotic transmission of macaque simplex virus 1 (MHV-1), also known as the herpes B virus, from macaques to humans (20). Although the virus causes little to no disease in its cognate host species, the virus spreads to the central nervous system in humans and ultimately causes permanent neurological deficits or even death. In persons who develop encephalomyelitis, the mortality rate is estimated to be about 80%

without antiviral treatment, illustrating the potential zoonotic threat that primate simplex viruses can cause to human health (85). More recently, a study revealed that the acquisition of herpes simplex virus 2 in a human ancestor was most likely the result of a cross-species transmission event of a chimpanzee herpes simplex virus (84). Conversely, infection of non-human primate species with HSV-1 can manifest in severe illness, often times resulting in death (86-88).

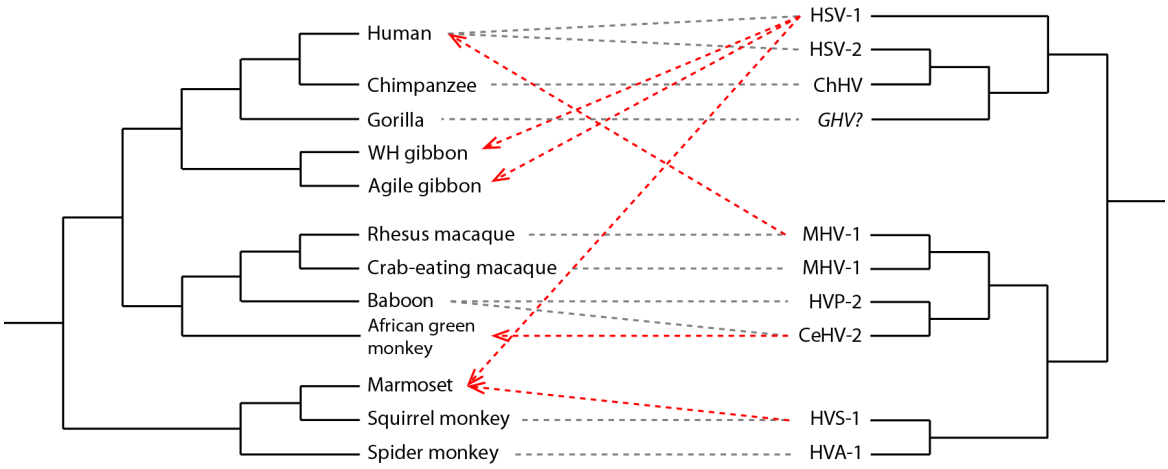


Figure 4-1: Primate simplex viruses and their natural hosts. The phylogenetic relationship of primate simplex viruses exactly mirrors that of the primate hosts (grey dotted lines), a finding that is consistent with codivergence. However, cross-species transmission events have been documented to occur in nature (red dotted lines), resulting in severe disease and often times death. All viruses shown in this figure have been isolated and sequenced with the exception of the gorilla herpes virus.

Interestingly, the large discrepancy in virulence of a single virus can be attributed to a small number of genetic differences between the host and non-host species it is able to infect (22). For example, the human genome is remarkably similar to that of the rhesus macaque genome, exhibiting 93% sequence identity (89). This means that less than 7% of the genome dictates whether MHV-1 can successfully infect humans and the severity of clinical disease that follows. Within this 7% are host genes that play critical roles in the

viral lifecycle and are the key determinants of cross-species transmission events. Therefore, to better understand the dynamics of cross-species transmission of viruses, it is of considerable interest to identify these host genes and if genetic differences between host and non-host species have a differential impact on viral replication.

For HSV-1, genes involved in the DNA damage response are potential candidates that may be crucial in modulating its pathogenicity. Amazingly, HSV-1 has evolved the ability to selectively activate specific arms of the DNA damage response that enhance its replication (65, 66, 90-92), while disabling those that are counterproductive to infection (67, 68, 93-97). In particular, the Mre11-Rad50-Nbs1 (MRN) complex, an early sensor of DNA damage, localizes to viral replication centers in actively infected cells (65, 90, 91). It has been shown that the presence of Mre11 drastically increases viral DNA replication and virus production, implicating the MRN complex as a cofactor in the HSV-1 lifecycle (65). In addition, several groups have reported an interaction between the MRN complex and HSV-1 proteins. A proteomic screen of host proteins that co-precipitated with the viral single-stranded DNA binding protein ICP8 revealed a potential interaction with Mre11 and Rad50 (98). The MRN complex has also been shown to physically interact with the HSV-1 exonuclease UL12 (99), while a peptide fragment of the multifunctional ICP0 protein was able to bind Nbs1 (97).

Previously, we reported that Nbs1 and several other DNA repair proteins have been the subject of intense positive selection in primates, with each primate species encoding different variants (11, 100). This is extremely unusual because rapidly evolving genes are known to be exclusively involved in environmental perception, sexual selection, and immunity (53, 54). In particular, because of the intimate relationship between HSV-1 and the MRN complex, we hypothesized that specific amino acid differences between the primate Nbs1 proteins could differentially affect the HSV-1

lifecycle. In this study, we tested the ability of several primate orthologs of Nbs1 to support HSV-1 replication and examined the role that Nbs1 plays in the viral lifecycle.

MATERIALS AND METHODS

Cell lines

NBS-ILB1 cells were grown in DMEM supplemented with 10% FBS, 100 μg/ml streptomycin, and 100 U/ml penicillin at 37°C and 5% CO_2. Cell lines stably expressing primate Nbs1 proteins with a c-terminal FLAG tag were generated using a retroviral transduction system and maintained under selection with media containing 800 μg/ml of G418 for at least 1 month. Expression of Nbs1 proteins was detected by Western blot using the Nbs1 specific antibody, Y112 (Genetex). Vero, U2OS, HEK-293, HEK-293Ts, and MDCK cells were purchased from ATCC and cultured in DMEM with 10% FBS and Penn/Strep.

Antibodies

Primary antibodies were purchased from Genetex (Nbs1 Y112, Mre11 12D7), Santa Cruz (actin, Rad50), Sigma (FLAG M2), and Abcam (GFP). The anti-Udorn antibody was a kind gift from Dr. Robert M. Krug. All secondary antibodies were purchased from Thermo Scientific.

Viruses

HSV-1

The HSV-1 strain 17 in1863 variant containing the *lacZ* gene in the thymidine kinase region was a kind gift from Dr. Chris Preston. The dl1043 virus has a 2 kb deletion in both copies of ICP0 and was obtained from Dr. Matthew Weitzman (101). The FXE virus encodes for ICP0 with a deletion in the RING domain (amino acids 106-149) and was also obtained from Dr. Matthew Weitzman (102). All virus stocks were grown on Vero cells. The in1863 strain was titered on either Veros or U2OS cells while the ICP0 mutant viruses (dl1043 and FXE) were titered on U2OS cells. Infections were conducted on cell monolayers at the indicated MOIs in serum free media for 1 hour at 37°C. The cells were washed with PBS and complete media containing 10% FBS and Pen/Strep was added. Unless otherwise noted, supernatant from infected cells were collected at the indicated timepoints.

For plaque assays, 10-fold serial dilutions were made and titered on the relevant cell lines as described above. After 1 hour of adsorption, media containing 10% FBS and 1% human serum was added. Plaques were stained using a 0.5% crystal violet, 25% methanol solution after 2-3 days. For plaque assays using the in1863 virus, cells were fixed with 2% paraformaldehyde and stained with an X-gal staining solution.

Adenovirus

Wild-type adenovirus serotype 5, recombinant adenoviral vectors encoding E1b55k and E4orf6, and the E4-deleted dl1004 virus were obtained from Matthew Weitzman. Cells were infected with adenovirus at the indicated MOI for 2 hours in DMEM with 2% FBS. After adsorption, the media was replaced with DMEM containing 10% FBS and Pen/Strep. Supernatants were harvested at the indicated timepoints.

Titering was performed on HEK-293 cells using the Adeno-X Rapid Titer Kit (Clontech). Whole cell lysates were harvested after 96 hours post infection and subjected to immunoblotting as described below.

Influenza

Single-cycle and multiple cycle influenza A virus infections (A/Udorn/H3N2; a kind gift from Robert M. Krug) were carried out at 2 MOI. Cells were washed with PBS and then incubated in infection media (DMEM supplemented with Pen/Strep, L-Glut, and 1% BSA) for one hour at 37°C. Cells were washed once more in PBS and incubated in influenza growth media (DMEM supplemented with Pen/Strep and L-Glut; multiple-cycle samples also included 0.5ug/ml N-acetylated trypsin). For single-cycle infections, cell lysates were harvested at 8 hours post infection using RIPA buffer supplemented with complete protease inhibitor (Roche) and PMSF (Invitrogen). Whole cell lysates were subjected to western blotting and influenza proteins were visualized using a polyclonal anti-Udorn antibody. For multiple-cycle infections, each cell line was seeded in triplicate and infected with a unique dilution of influenza A virus. Supernatants were collected at 1, 12, 24, and 48 hours post infection and titered by carrying out a plaque assay on MDCK reporter cells. Infections for plaque assays were carried out as described above, except the final incubation media contained 1.4% Avicel (Sigma) to induce plaque formation.

Immunoprecipitation and immunoblotting

MRN co-IP

NBS-ILB1 cells expressing empty vector, human, or white-cheeked gibbon Nbs1 were lysed in ice-cold lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl, 0.5% NP-40) and

rotated for at least 1 hour at 4°C. The lysates were cleared and a small aliquot was saved as the input sample. The remaining samples were incubated with 10 µl of anti-DYKDDDDK conjugated magnetic beads (Syd labs) at 4°C for at least 2 hours. The beads were washed 3 times with lysis buffer and 3 times with Buffer A (25 mM Tris pH 8, 100 mM NaCl, 10% v/v glycerol, 1 mM DTT). Bound proteins were eluted with a DYKDDDDK peptide. SDS loading buffer was added to the samples and boiled for 10 minutes. The samples were separated via electrophoresis on a polyacrylamide gel and transferred to a nitrocellulose membrane. The membrane was blocked in 5% milk in TBS-T and immunoblotting was carried out using a primary antibody and the appropriate HRP-conjugated secondary antibody. Amersham ECL Prime Western Blotting Detection Reagent (GE Life Sciences) was used for visualization.

ICP0 and Nbs1 co-IP

HEK-293Ts were transfected with the indicated plasmids using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. At 48 hours post transfection, cells were harvested in 500 µl of ice-cold co-IP buffer with protease inhibitors (50 mM Tris-HCl pH7.4, 150 mM NaCl, 0.1% Triton X-100, 50 mM NaF, 1 mM sodium vanadate) and subjected to mild sonication. The lysates were cleared and a small aliquot was saved as the input sample. The remaining samples were incubated with the anti-GFP antibody (Abcam) for at least 2 hours at 4°C with constant rotation. Dynabeads Protein G (Novex) were then added to the samples and rotated at 4°C for at least 1 hour. The beads were washed four times with ice-cold co-IP buffer and resuspended in SDS loading buffer. Immunoblotting was performed as described in the previous section.

DNA repair assays

Cells were plated at a density of 200 cells per well in 6 well plates. The following day, media containing the appropriate concentrations of either camptothecin (Sigma) or hydroxyurea (Sigma) were added to the cells. 24 hours later, the media was replaced with fresh media. For X-ray irradiation experiments, the cells were subjected to the doses indicated using the Faxitron X-ray system at 120 kV and 5mA. All cells were grown to allow for colony formation at 37°C for at least 7 days. Colonies were stained with a crystal violet staining solution, washed, and counted. Cell counts from treated wells were normalized to untreated controls and expressed as percent survival.

Viral DNA detection

Cells were infected with HSV-1 or adenovirus and collected at the indicated timepoints. Viral DNA was extracted from cells using the DNeasy Blood and Tissue Kit (QIAGEN) and quantified by amplifying the ICP27 gene for HSV-1 and DBP for adenovirus. These values were normalized to an endogenous control gene such as RPLP0 or tubulin and then to the 4 hour input sample. Products were amplified using the Power SYBR Green PCR Master Mix (Applied Biosystems) on the ViiA7 Real-Time PCR System (Applied Biosystems).

RESULTS

Species-specific effects of Nbs1 on HSV-1 replication

While NBS1 has been rapidly evolving over the course of primate speciation, the genes encoding the remaining members of the MRN complex, RAD50 and MRE11, do not display the same pattern (11). In fact, the Mre11 and Rad50 proteins encoded by

orangutan, white-cheeked gibbon, and rhesus macaque are greater than 99% identical in amino acid composition to the human proteins, while the Nbs1 proteins encoded by these species show greater levels of divergence from the human Nbs1 (Figure 4-2A). These non-human primate Nbs1 proteins differ by 21 to 36 amino acids from the human homolog. Furthermore, these differences are scattered throughout the length of the protein (Figure 4-2B). It is interesting to note that key regions of the protein known to mediate critical protein-protein interactions with other DNA repair factors, such as Mre11 and ATM, are essentially conserved in all primate orthologs of Nbs1 (alignment shown in Figure 4-3).

A.

% protein identity

	Mre11	Rad50	Nbs1	ΔAAs
Human	--	--	--	--
Orangutan	>99%	>99%	97.2%	21
WC Gibbon	>99%	>99%	97.2%	21
Rhesus	>99%	>99%	95.2%	36

B.

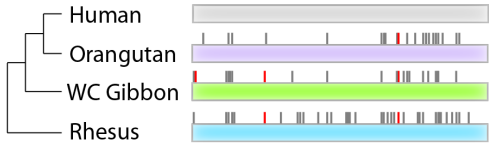


Figure 4-2: Nbs1 exhibits the greatest sequence divergence in primates. (A) The human Mre11, Rad50, or Nbs1 amino acid sequence was aligned to the orangutan, white-cheeked gibbon, or rhesus macaque protein sequence and the percent protein identity was calculated. **(B)** Amino acid differences in the primate Nbs1 orthologs when compared to human Nbs1. Tick marks represent positions in which the primate Nbs1 ortholog encodes a different amino acid than the human Nbs1 protein. Red tick marks represent sites under positive selection as described previously (11).

Although the role of Nbs1 has never been directly characterized in the HSV-1 lifecycle, previous work has shown that HSV-1 titers are reduced by greater than 10 fold in the absence of Mre11, suggesting that either Mre11 or the MRN complex acts as a cofactor in the HSV-1 lifecycle (65). To test whether Nbs1 also has a positive effect on

the HSV-1 lifecycle, we obtained a cell line hypomorphic for Nbs1, NBS-ILB1 (103), and stably complemented it with the human NBS1 allele or the empty vector (ΔNBS1). Cells were infected with a wild-type HSV-1 strain encoding the lacZ gene at a multiplicity of infection (MOI) of 0.01 and virus production was measured (Figure 4-4A). At 24 hours post infection, viral titers in the supernatant were significantly increased in cells expressing human Nbs1 (red bar) when compared to the parental cell line (ΔNBS1, blue bar). This result is complementary to previous studies showing an effect of Mre11 on HSV-1 production and supports a model where the MRN complex aids the HSV-1 lifecycle.

Next, we tested whether Nbs1 homologs encoded by other non-human primate species would similarly enhance virus yields in these Nbs1-deficient cells. NBS-ILB1 cells were transduced to stably express the Nbs1 homologs of different non-human primate species. The orangutan (purple bar) and rhesus macaque (cyan bar) Nbs1 also supported higher levels of HSV-1 production, although not to the extent of human Nbs1 (Figure 4-4A). Surprisingly, the white-cheeked gibbon allele (green bar) exhibited minimal to no gains in virus production, with viral titers equaling that of the parental, Nbs1-deficient cell line. This differential effect on HSV-1 production was not attributed to Nbs1 expression levels (Figure 4-4B, inset).

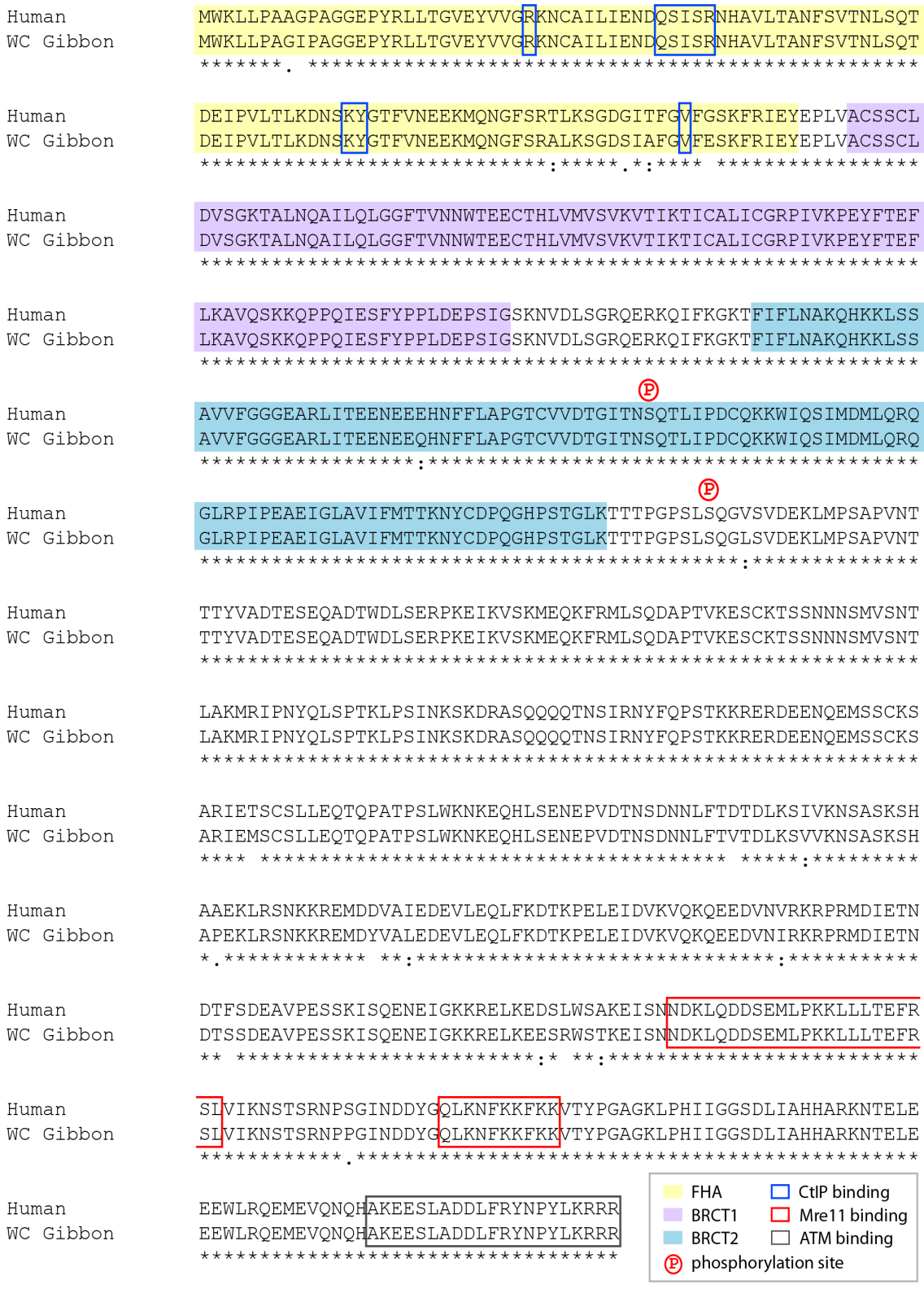

Human MWKLLPAAGPAGGEPYRLLTGVEYVVGRKNCAILIENDQSISRNHAVLTANFSVTNLSQT
WC Gibbon MWKLLPAGIPAGGEPYRLLTGVEYVVGRKNCAILIENDQSISRNHAVLTANFSVTNLSQT
Human DEIPVLTLKDNSKYGTFVNEEKMQNGFSRTLKSGDGITFGVFGSKFRIEYEPLVACSSCL
WC Gibbon DEIPVLTLKDNSKYGTFVNEEKMQNGFSRALKSGDSIAFGVFESKFRIEYEPLVACSSCL
Human DVSGKTALNQAILQLGGFTVNNWTEECTHLVMVSVKVTIKTICALICGRPIVKPEYFTEF
WC Gibbon DVSGKTALNQAILQLGGFTVNNWTEECTHLVMVSVKVTIKTICALICGRPIVKPEYFTEF
Human LKAVQSKKQPPQIESFYPPLDEPSIGSKNVDLSGRQERKQIFKGKTFIFLNAKQHKKLSS
WC Gibbon LKAVQSKKQPPQIESFYPPLDEPSIGSKNVDLSGRQERKQIFKGKTFIFLNAKQHKKLSS
Human AVVFGGGEARLITEENEEEHNFFLAPGTCVVDTGITNSQTLIPDCQKKWIQSIMDMLQRQ
WC Gibbon AVVFGGGEARLITEENEEQHNFFLAPGTCVVDTGITNSQTLIPDCQKKWIQSIMDMLQRQ
Human GLRPIPEAEIGLAVIFMTTKNYCDPQGHPSTGLKTTTPGPSLSQGVSVDEKLMPSAPVNT
WC Gibbon GLRPIPEAEIGLAVIFMTTKNYCDPQGHPSTGLKTTTPGPSLSQGLSVDEKLMPSAPVNT
Human TTYVADTESEQADTWDLSERPKEIKVSKMEQKFRMLSQDAPTVKESCKTSSNNNSMVSNT
WC Gibbon TTYVADTESEQADTWDLSERPKEIKVSKMEQKFRMLSQDAPTVKESCKTSSNNNSMVSNT
Human LAKMRIPNYQLSPTKLPSINKSKDRASQQQQTNSIRNYFQPSTKKRERDEENQEMSSCKS
WC Gibbon LAKMRIPNYQLSPTKLPSINKSKDRASQQQQTNSIRNYFQPSTKKRERDEENQEMSSCKS
Human ARIETSCSLLEQTQPATPSLWKNKEQHLSENEPVDTNSDNNLFTDTDLKSIVKNSASKSH
WC Gibbon ARIEMSCSLLEQTQPATPSLWKNKEQHLSENEPVDTNSDNNLFTVTDLKSVVKNSASKSH
Human AAEKLRSNKKREMDDVAIEDEVLEQLFKDTKPELEIDVKVQKQEEDVNVRKRPRMDIETN
WC Gibbon APEKLRSNKKREMDYVALEDEVLEQLFKDTKPELEIDVKVQKQEEDVNIRKRPRMDIETN
Human DTFSDEAVPESSKISQENEIGKKRELKEDSLWSAKEISNNDKLQDDSEMLPKKLLLTEFR
WC Gibbon DTSSDEAVPESSKISQENEIGKKRELKEESRWSTKEISNNDKLQDDSEMLPKKLLLTEFR
Human SLVIKNSTSRNPSGINDDYGQLKNFKKFKKVTYPGAGKLPHIIGGSDLIAHHARKNTELE
WC Gibbon SLVIKNSTSRNPPGINDDYGQLKNFKKFKKVTYPGAGKLPHIIGGSDLIAHHARKNTELE
Human EEWLRQEMEVQNQHAKEESLADDLFRYNPYLKRRR
WC Gibbon EEWLRQEMEVQNQHAKEESLADDLFRYNPYLKRRR
FHA
BRCT1
BRCT2
CtIP binding
Mre11 binding
ATM binding
phosphorylation site

Figure 4-3: Protein sequence alignment of human and white-cheeked gibbon Nbs1. A protein sequence alignment was generated using ClustalX. The FHA, BRCT1, and BRCT2 domains are shown in yellow, purple, and blue, respectively (104). Sites of phosphorylation are indicated in red. Blue open boxes highlight residues important for binding CtIP (105), red boxes are residues that mediate interaction with Mre11 (106), and grey boxes are amino acids involved in ATM binding (107).

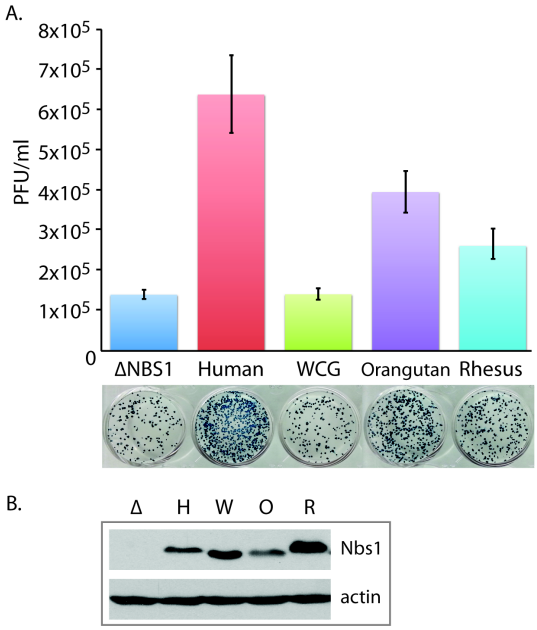


Figure 4-4: Primate orthologs of Nbs1 enhance HSV-1 production to varying degrees. (**A**) NBS-ILB1 cells complemented with empty vector, human, white-cheeked gibbon (WCG), orangutan, or rhesus macaque Nbs1 were infected with the in1863 HSV-1 strain at an MOI of 0.01. Virus in the supernatant was harvested at 24 hours post infection and titered on Vero cells. The resulting plaques were counted (inset) and viral titers were calculated. Results shown are an average of three independent replicates with error bars representing standard deviations. (**B**) Complemented NBS-ILB1 cells were harvested and Western blot analysis was conducted to determine the expression level of Nbs1 in each of the cell lines. Actin was used as a loading control. Δ – empty vector, H – human, W – white-cheeked gibbon, O – orangutan, R – rhesus macaque.

To explore this further, the ΔNBS1, human, and white-cheeked gibbon Nbs1 complemented cell lines were infected at MOI 0.01 and progeny viruses were collected at 12, 24, 36, and 48 hours post infection (Figure 4-5A). The measured viral yields from cells expressing human Nbs1 was consistently elevated at 24, 36, and 48 hours when compared to titers produced in the parental cell line (ΔNBS1), up to 37 fold higher at 24 hours post infection. Interestingly, the viral titers from white-cheeked gibbon Nbs1 expressing cells were essentially identical to that of the ΔNBS1 cell line at all timepoints analyzed. Collectively, these results indicate that only a small number of amino acids in

Nbs1 dictate the extent to which Nbs1 can augment HSV-1 production. Human and white-cheeked gibbon NbS1 differ by only 21 amino acids (Figure 4-2A), yet human Nbs1 supports virus infection, while white-cheeked gibbon Nbs1 has a phenotype equivalent to the Nbs1-deficient parental line (Figure 4-5A). This is despite the fact both of these homologs are wildtype alleles, presumably supporting full DNA repair capabilities within their respective species.

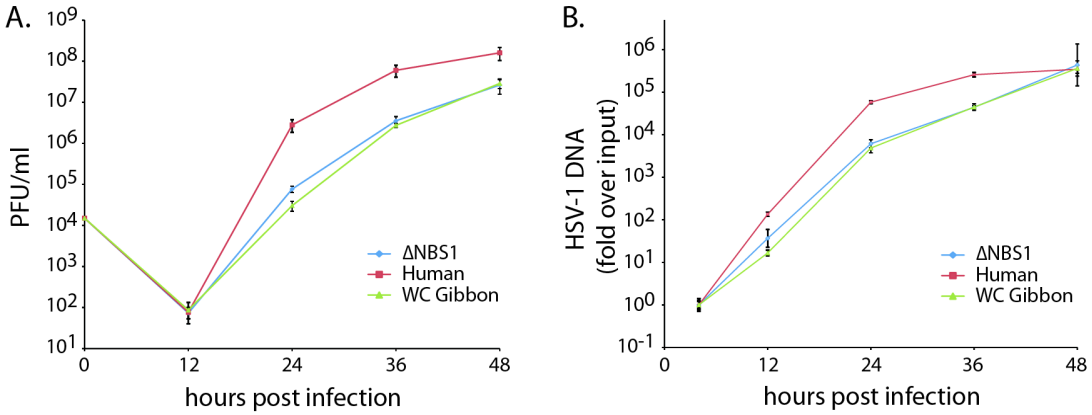


Figure 4-5: HSV-1 DNA replication is affected by Nbs1. (**A**) NBS-ILB1 cells complemented with empty vector, human, or white-cheeked gibbon Nbs1 were infected with the in1863 HSV-1 strain at an MOI of 0.01. Supernatant containing newly produced viruses were collected at 12, 24, 36, and 48 hours post infection and viral titers were determined by plaque assay on Vero cells. Results are presented as an average of three replicates with error bars representing standard deviations. (**B**) Infected cells in **A** were collected and DNA was extracted. The amount of viral DNA present in each of the samples was determined by qPCR with *ICP27* primers and normalized to the endogenous control gene *RPLP0*. The amount of viral DNA is expressed as a fold increase relative to the 4 hour timepoint (input) and error bars represent standard error.

The MRN complex is thought to be important for assisting in the replication of the DNA genome of HSV-1 (65). To determine if this surprising effect on virus production was consistent with this model, DNA was extracted from cells at several timepoints after infection (cells from Figure 4-5A) and viral DNA was quantified by

qPCR. As shown in Figure 4-5B, human Nbs1 supported viral DNA replication to a greater extent than the white-cheeked gibbon Nbs1 or the empty vector at 12, 24, and 36 hours post infection. However, at 48 hours, all cells had approximately equal amounts of viral DNA present. In summary, we have shown a species-specific ability of wildtype Nbs1 orthologs to support the replication of human HSV-1.

DNA repair functions of the white-cheeked gibbon Nbs1 are intact in human cells

The protein sequences of human and white-cheeked gibbon Nbs1 differ at only 21 amino acid positions. As mentioned previously, there is a paucity of amino acid changes in regions that are known to be important for function, such as the Mre11 binding domain, ATM binding region, and those that mediate CtIP binding (Figure 4-3). Nonetheless, to exclude the possibility that the white-cheeked gibbon Nbs1 is nonfunctional in human cells, we carried out a number of experiments.

First, co-immunoprecipitation was performed to test the interaction of white-cheeked gibbon Nbs1 with human Mre11 and Rad50. Both the human and white-cheeked gibbon Nbs1 were able to interact with human Mre11 and Rad50 in proportion to Nbs1 expression levels (Figure 4-6). In contrast, no Nbs1 was detected in ΔNBS1 cells and consequently, Mre11 and Rad50 could not be detected in the immunoprecipitated samples.

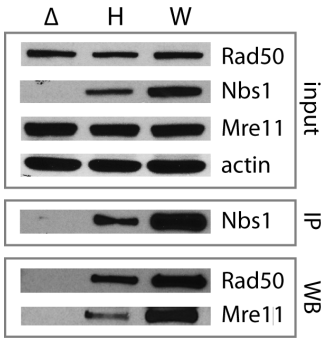


Figure 4-6: Formation of the MRN complex is conserved. NBS-ILB1 cells complemented with the indicated *NBS1* allele or empty vector (Δ) were lysed and immunoprecipitated using an Nbs1 antibody. The ability of human (H) and white-cheeked gibbon (W) Nbs1 to form a complex with human Mre11 and Rad50 was assessed by immunoblotting.

Next, we tested the integrity of DNA repair in each of these cell lines by measuring sensitivity to three different types of genotoxic stress: camptothecin, ionizing radiation, and hydroxyurea. Decreased cell viability is characteristic of hypersensitivity to a particular DNA damaging agent and can be assessed by a colony formation assay. When exposed to increasing levels of camptothecin or X-ray irradiation, both the human and white-cheeked gibbon Nbs1 were able to complement the deficiency in DNA repair activity when compared to cells that do not express Nbs1 (Figure 4-7A and 4-7B, respectively). No significant differences between any of the three cell lines were observed when hydroxyurea was used, although survival of cells expressing the white-cheeked gibbon or human Nbs1 was slightly higher than Nbs1-deficient cells (Figure 4-7C). Collectively, these data suggest that the functionality of the white-cheeked gibbon Nbs1 is preserved in human cells, and that the inability to increase HSV-1 titers is not simply due to a defect in its DNA repair functions in the context of a human cell.

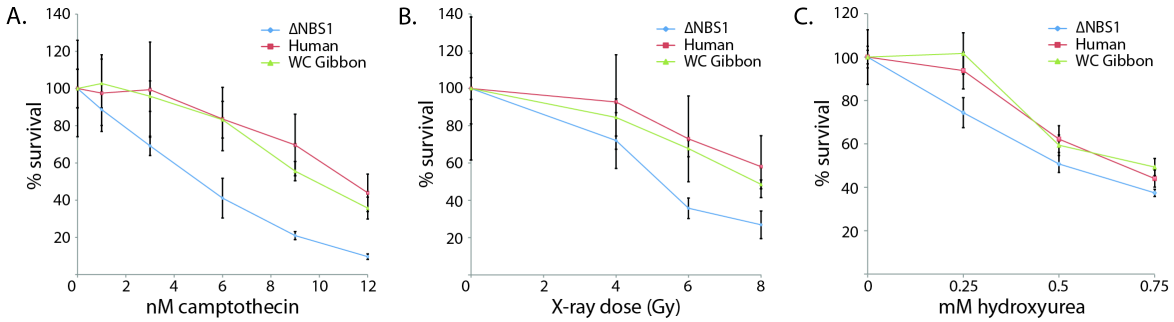


Figure 4-7: DNA repair activity of Nbs1 is conserved among primates. NBS-ILB1 cells expressing the empty vector, human, or white-cheeked gibbon Nbs1 were exposed to increasing doses of (**A**) camptothecin, (**B**) X-rays, and (**C**) hydroxyurea. Colony formation was assessed after 7-10 days. Percent survival was calculated by normalizing the number of colonies to untreated controls. Experiments were performed in triplicate and error bars represent standard deviations from the mean.

Nbs1 does not affect the lifecycle of adenovirus and influenza

Because the 21 amino acid differences between human and white-cheeked gibbon Nbs1 specifically affect HSV-1 replication and not the housekeeping functions of Nbs1 in the cell, we wanted to next test whether the replication of other viruses that gain nuclear access could be affected. Adenovirus is another DNA virus that replicates in the nucleus and has been shown to extensively interact with DNA repair machinery. Specifically, the MRN complex orchestrates the concatenation of newly synthesized adenoviral genomes, thus preventing further DNA replication and packaging into viral particles (64). In turn, adenovirus encodes proteins that counteract the antiviral effects of the host MRN complex. The viral E4orf3 protein mislocalizes the MRN complex into nuclear tracks and aggresomes (64, 108-111), while the E4orf6/E1b55k complex is responsible for degrading MRN and other DNA repair constituents (64, 112-114).

To determine if the different Nbs1 proteins were susceptible to degradation by adenoviral proteins, we infected our stably complemented cell lines with recombinant

adenoviral vectors encoding the Ad5 E4orf6 and E1b55k proteins (Figure 4-8A). Both the human and white-cheeked gibbon Nbs1 were degraded in the presence of the E4orf6/E1b55k complex, along with other known cellular targets such as Rad50 and p53. Degradation of the MRN complex inhibits concatemer formation, and therefore virus production was not expected to differ among the cell lines. Indeed, when cells were infected with wild-type Ad5, no differences in viral yield were observed at 24 and 48 hours post infection (Figure 4-8B). Thus, the species-specific differences seen in Nbs1 do not have an impact on adenoviral infection and production.

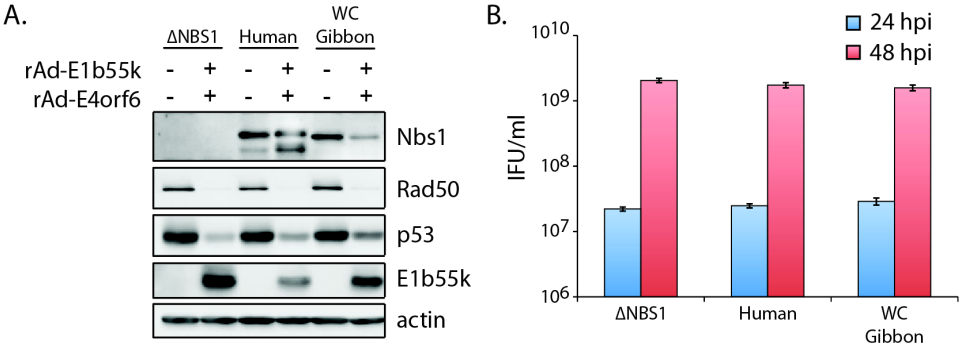


Figure 4-8: Sequence variation in Nbs1 does not affect the adenoviral lifecycle. (**A**) NBS-ILB1 cells expressing empty vector, human, or white-cheeked gibbon Nbs1 were infected with recombinant adenoviral vectors encoding the adenoviral E1b55k and E4orf6 proteins. Cells were harvested 96 hours post infection to detect degradation of the Nbs1 and other proteins targeted by the viral complex. Experiment performed by Neha Pancholi. (**B**) Cells were infected with wild-type adenovirus in triplicate. 24 and 48 hours post-infection, supernatants were collected and titered on HEK-293 cells. The results are expressed as an average of three replicates and error bars represent standard deviations.

As mentioned above, in the absence of the viral E4 genes, the host MRN complex mediates the concatenation of adenoviral genomes and hinders viral genome replication. We were interested to see if both the human and white-cheeked gibbon Nbs1 proteins were equally capable of recognizing foreign adenoviral DNA and mounting this unique

host defense mechanism. We next infected our stably complemented cell lines with a mutant adenovirus lacking the E4 coding region (dl1004 Ad5). In cells expressing human Nbs1, a 10 fold decrease in the amount of viral DNA was apparent when compared to cells deficient in Nbs1 (Figure 4-9). Viral DNA replication was also reduced to the same extent in the presence of white-cheeked gibbon Nbs1. Collectively, these results show that both Nbs1 orthologs equally recognize actively replicating adenoviral genomes and are equally susceptible to the anti-MRN antagonists encoded by the virus.

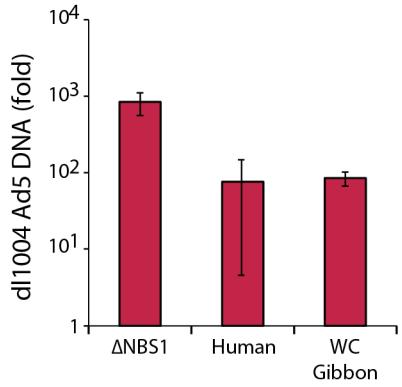


Figure 4-9: Human and white-cheeked gibbon Nbs1 are antiviral in the absence of the viral E4 genes. (**A**) NBS cells were infected with an E4-deleted adenovirus 5. After 4 and 30 hours post-infection, cells were harvested and DNA was extracted. qPCR was performed using primers that amplify the adenoviral DBP gene and normalized to the endogenous tubulin control gene. These values were then expressed relative to the 4 hour input sample to obtain the values shown. The results are expressed as an average of three biological replicates and error bars represent standard deviations. Experiment performed by Neha Pancholi.

Influenza virus also gains access to the nucleus during its lifecycle. Replication of this RNA virus does not involve DNA intermediates and DNA repair responses are not expected to have an effect on influenza infection. First, the complemented cells were infected with the Udorn H3N2 influenza strain at MOI 2 and cell lysates were collected at 12 hours post infection (Figure 4-10A). Expression of the viral proteins hemagglutinin,

nucleoprotein, and matrix were consistent across all cell lines tested, indicating that flu protein production is not affected by Nbs1. In addition, we were not able to discern any differences in the kinetics of virus production between the cell lines at several timepoints post infection (Figure 4-10B). Collectively, these data show that Nbs1 does not play a role in the influenza lifecycle and further supports our conclusion that species-specific differences in Nbs1 specifically affect HSV-1 and not general cell function or interactions with other viruses.

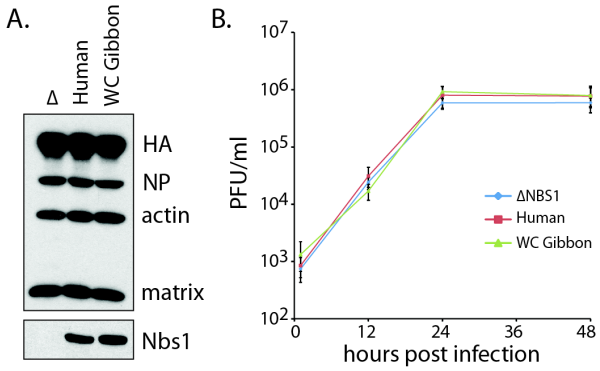


Figure 4-10: Sequence variation in Nbs1 does not affect the influenza lifecycle. (**A**) Cells were infected with the A/Udorn/H3N2 influenza A virus at an MOI of 2. At 12 hours post infection, cells were harvested and lysates were subjected to western blotting. Influenza proteins were visualized using an anti-Udorn antibody that recognizes hemagluttinin (HA), nucleoprotein (NP), and matrix (M). Nbs1 and actin expression levels were also determined. (**B**) Cells were infected with influenza as described in **A**. Supernatants were collected at 1, 12, 24, and 48 hours post infection and titers were determined by plaque assay on MDCK cells.

Interaction between ICP0 and Nbs1 is species-specific

In the absence of Mre11, an ICP0 null HSV-1 strain displays severely crippled replication kinetics when compared to Mre11 complemented cells (65). The difference in virus yield between complemented and deficient cells is similar to wild-type virus

kinetics. This suggests that Mre11's involvement in the HSV-1 lifecycle is independent of ICP0 functions. Furthermore, an ICP0 peptide fragment has been shown to bind human Nbs1 but not Mre11 (97), potentially signifying that Nbs1 may participate in HSV-1 replication in a manner that is complementary but independent of Mre11.

First, we set out to determine the viral growth kinetics of the ICP0 null virus in our Nbs1 complemented cell lines. Unlike the patterns seen with the wild-type virus, the presence of human Nbs1 only increased virus titers about 3 fold at 36 and 48 hours post infection (Figure 4-11A). By 60 hours, no discernable differences in virus production were seen between the Nbs1 deficient and human Nbs1 complemented cells. A virus with a deletion in RING domain of ICP0 displayed growth kinetics similar to wild-type HSV-1, indicating that ICP0 E3 ubiquitin ligase activity is not responsible for the ability of HSV-1 to hijack human Nbs1 function (Figure 4-11B). Unlike the previous reported effect of Mre11 on ICP0 null virus titers, our data indicate that a yet unidentified interaction could exist between ICP0 and human Nbs1.

As mentioned previously, a peptide fragment of ICP0 has been shown to interact with the human Nbs1 protein. We next wished to recapitulate this interaction between full-length ICP0 and human Nbs1 and also to determine if the white-cheeked gibbon protein was capable of interacting with ICP0. 293Ts were transfected with plasmids encoding ICP0-GFP and FLAG-tagged human Nbs1 or white-cheeked gibbon Nbs1. ICP0-GFP was immunoprecipitated with a GFP antibody followed by immunoblotting with an anti-FLAG and anti-GFP antibody. As shown in Figure 4-12, we were able to confirm that human Nbs1 did indeed interact with ICP0 (lane 2). Conversely, we were surprised to see a lack of interaction between white-cheeked gibbon Nbs1 and ICP0 (lane 4), a finding that strongly correlates to the decreased viral production phenotype observed

above (Figure 4-5A). In all cases, no detectable Nbs1 was precipitated in the absence of ICP0 (lanes 1 and 3).

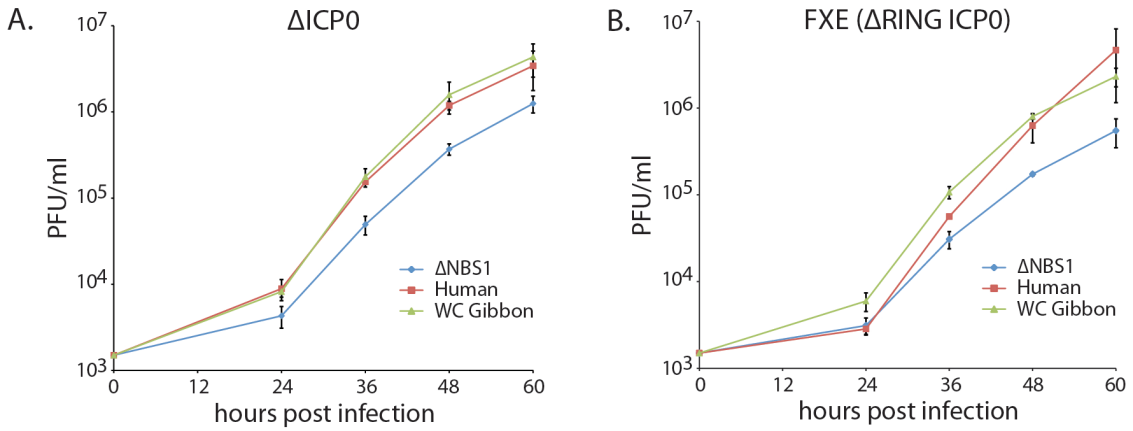


Figure 4-11: Nbs1 potentially interacts with ICP0. NBS cells expressing the empty vector, human, or white-cheeked gibbon Nbs1 were infected with a virus containing a deletion in the ICP0 coding regions (ΔICP0, **A**) or an ICP0 RING domain mutant virus (FXE, **B**) at an MOI of 0.3. Supernatants were collected at 24, 36, 48, and 60 hours post infection. Viruses were titered U2OS cells. The titers shown are an average of three independent replicates, with error bars representing standard deviations.

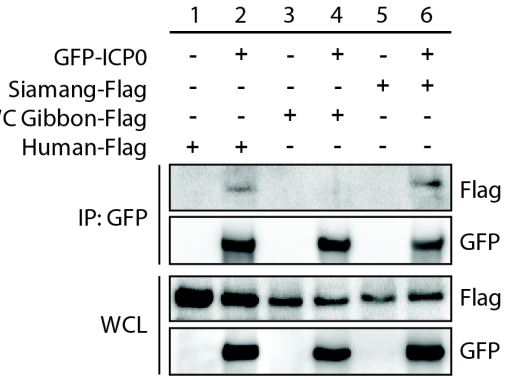


Figure 4-12: Nbs1 interacts with ICP0 in a species-specific manner. HEK-293T cells were transfected with the indicated plasmids. Immunoprecipitation was performed using an anti-GFP antibody to precipitate ICP0-GFP and immunoblotting was performed to detect Nbs1 binding. Experiment performed by Eui Tae Kim.

To further determine if this lack of interaction correlated with viral titers, we examined the ability of Nbs1 from a closely related gibbon species to bind ICP0. The Nbs1 protein of siamang differs from the white-cheeked gibbon variant by only 12 amino acids. However, siamang Nbs1 expression greatly enhances virus yield in a manner similar to the human protein (Figure 4-13) and is able to co-precipitate with ICP0 (Figure 4-12), further supporting our findings. This data, in combination with the wild-type virus production results, implies that ICP0 may bind and recruit human and siamang Nbs1, but not the white-cheeked gibbon Nbs1 protein, for efficient virus replication. Therefore, species-specific differences in Nbs1 may be an important determinant for maximal HSV-1 infection.

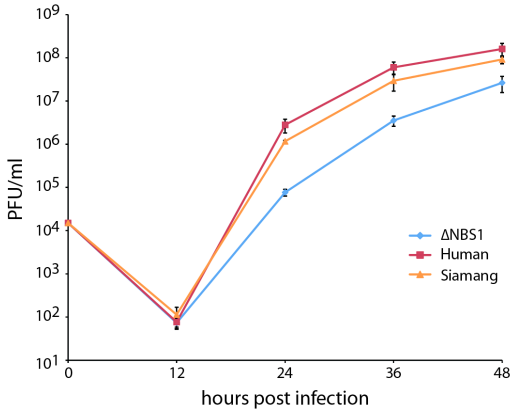


Figure 4-13: Siamang Nbs1 supports HSV-1 replication in a manner similar to human Nbs1. NBS-ILB1 cells complemented with human or siamang Nbs1 were infected with HSV-1 at an MOI of 0.01. Supernatant was collected at 12, 24, 36, and 48 hours post-infection and titered on Veros. The results shown are an average of three replicates with error bars representing standard deviations.

Specific residues in Nbs1 are important for HSV-1 replication

Siamangs and white-cheeked gibbons are lesser apes that are separated by six to seven million years of divergence (115). As mentioned above, only 12 out of the 754

residues differ between the Nbs1s encoded by these two species, most of which are concentrated towards the C-terminus of the protein. To determine if specific amino acids in Nbs1 are responsible for the differences in supporting HSV-1 replication, we replaced four amino acids in the white-cheeked gibbon Nbs1 with the residues from the siamang Nbs1 (Figure 4-14A, WS-1). The corresponding mutations were also made in the siamang allele (SW-2). We then generated stable cell lines expressing these constructs and infected them with HSV-1. Much to our surprise, WS-1 supported HSV-1 replication to the same degree as the siamang Nbs1 while SW-2 levels of virus production mirrored that of white-cheeked gibbon Nbs1 (Figure 4-14B). This result indicates that just four amino acids in Nbs1 (residues 603, 624, 631, and 673) are responsible for the large differences in virus titers produced in these cell lines. These differences were not attributable to a defect in DNA repair activity as assessed by camptothecin sensitivity (data not shown).

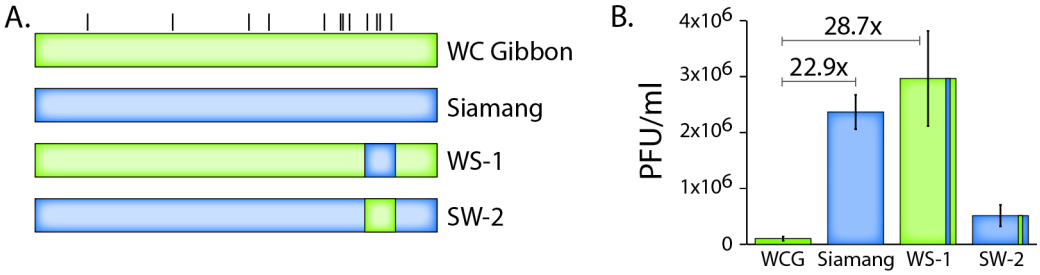


Figure 4-14: Key residues in Nbs1 are responsible for differences in virus production. (**A**) Chimeric gibbon Nbs1 constructs generated for interaction assays. (**B**) NBS1 cells expressing the indicated Nbs1 constructs were infected with wild-type HSV-1 at an MOI of 0.01. Viruses were harvested from the supernatant at 30 hours post infection and titered Veros. The results are an average of three independent replicates with error bars representing standard deviations from the mean.

DISCUSSION

We present for the first time that a host protein that is neither a receptor nor an antiviral restriction factor, may act as a barrier to cross-species transmission of HSV-1. Although viral replication does occur to some extent in cells lacking Nbs1 and in the presence of white-cheeked gibbon Nbs1, it is dramatically crippled when compared to viral titers produced in human Nbs1 and other primate Nbs1 complemented cells (Figure 4-5A). However, the MOIs used in this study are likely to be higher than naturally occurring HSV-1 infections, possibly pushing the balance towards active viral replication. Nonetheless, slower replication kinetics in the context of a whole organism could allow ample time for full activation of the immune response, potentially resulting in a profound impact on the virulence of a virus. In addition, non-permissive Nbs1 variants and other host restriction factors could potentially work in concert to further restrict HSV-1 from actively replicating in infected cells.

Our data clearly show that Nbs1 is co-opted by HSV-1 in a species-specific manner (Figure 4-4A). In particular, key residues in Nbs1 determine the degree to which this DNA repair protein can positively influence viral replication (Figure 4-14B). This effect is not due to differences in the intrinsic functions of the primate Nbs1 orthologs as assessed by sensitivity to genotoxic stress. This leaves the possibility that these residues mediate interaction with a viral protein that may recruit Nbs1 to enhance virus replication. It is intriguing that the ability of ICP0 to bind the different primate orthologs of Nbs1 differentially correlates to the drastic differences in virus production seen (Figure 4-12). Furthermore, this interaction does not facilitate degradation of Nbs1, unlike other host factors known to interact with ICP0 (93, 94, 97, 116-118). However, the RING domain that mediates ubiquitination and degradation of other cellular substrates may be important in hijacking Nbs1 functions (Figure 4-11B). How Nbs1 specifically affects

HSV-1 replication, and whether or not ICP0 directly interacts with or ubiquitinates Nbs1 is currently under investigation.

Additionally, this study has further revealed additional information on how HSV-1 hijacks the host DNA repair machinery to promote productive infection. Much like Mre11 (65), the presence of specific Nbs1 variants augment viral DNA replication and virus yield (Figure 4-5). At first glance, this is not surprising because Mre11 and Nbs1 exist in a complex within the cell. However, key pieces of evidence suggest that the roles that Mre11 and Nbs1 play in the HSV-1 lifecycle may be complimentary, but separable. For example, a proteomic analysis has shown that Mre11 and Rad50, but not Nbs1, interact with the viral single-stranded DNA binding protein ICP8 (98). In another study, an ICP0 peptide fragment was able to co-precipitate with Nbs1, but not other components of the MRN complex (97). Furthermore, although replication of the ICP0-null virus in our Nbs1-deficient cells displays slightly impaired growth kinetics when compared to Nbs1 complemented cells (Figure 4-11A), the disparity in viral titers is nowhere near the differences seen in Mre11-deficient and Mre11 complemented cell lines (65). Further studies are warranted to better understand the dynamic interactions between individual MRN components and HSV-1 proteins.

CONCLUSIONS

New diseases, including new human diseases, emerge when viruses evolve compatibility with a novel host species. For this reason, it is critical to define the host components that dictate viral host range. Here, we show that compatibility with a DNA repair factor, Nbs1, is a critical determinant of cross-species transmission for herpes simplex virus 1 (HSV-1) in primates. We show that the viral ICP0 protein of HSV-1

interacts with human Nbs1, and that Nbs1 facilitates genome replication of this DNA virus. We show that the ICP0 interaction with Nbs1 is species-specific, and does not occur with all primate homologs of Nbs1. These results broaden our current understanding of the host determinants of cross-species transmission to include essential, nuclear proteins like Nbs1.

Chapter 5: Concluding Remarks

REDUCING THE ERROR RATE OF HIGH-THROUGHPUT SEQUENCING

In chapter 2, I present a novel library preparation method with an accompanying bioinformatics pipeline that reduces the error rate of high-throughput sequencing data, allowing for the acquisition of high quality, low error reads. Although high-throughput sequencing technologies have expanded and revolutionized a variety of scientific fields, the error rates associated with these technologies do not allow for accurate measurements of variation within heterogenous samples. In these cases, true genetic variation within a sample cannot be distinguished from sequencing errors and hinders the study of complex samples, such as viral populations and mutations arising from inaccurate DNA repair. As error rates of high-throughput sequencing technologies continue to improve, the error rate problem may become more approachable, even without error-correction methods. However, even if error rates can be substantially decreased without the use of library preparation schemes, our method and other competing methods can be adapted to allow for more accurate "counting" of unique molecules.

Although we were able to achieve error rates that were equivalent to the gold standard of Sanger sequencing methods, there are additional modifications that are currently underway in order to further drive the error rates of our method down while still maintain high (or higher) efficiencies. Using the approach of the duplex barcoding method (7), we are developing a variation of our circle-sequencing method that utilizes the information from both strands of a duplex DNA molecule. Although mutations can occur at any step along the sample preparation process, they will do so at random along the length of the DNA molecule, and in most cases, do not occur in both bases of a

complementary base pair in the absence of DNA repair mechanisms. In fact, our lab already has evidence to suggest that this new approach does indeed reduce the error rate much more dramatically than duplex barcoding methods. Harnessing the strengths of duplex barcoding and circle sequencing into one method may drive the error rates down even further without sacrificing efficiency.

POSITIVE SELECTION IN BRCA1 AND BRCA2

In chapter 3, I describe the unique evolutionary trajectory seen in *BRCA1* and *BRCA2*. Specific residues have been identified to be rapidly evolving in response to a strong selective pressure that is yet to be uncovered. Our dataset, which is comprised of a large number of very closely primate species, shows that selection is acting upon humans and our closest living relatives. This allows us to speculate that a selective pressure specific to primates is driving the rapid evolution seen in both BRCA1 and BRCA2. Unlike the large datasets used by previous groups, which included widely divergent mammal species, our dataset allows us to show that these evolutionary signatures are strong enough to detect with high confidence within closely related species and that specific selective pressures may be at play within these species. One can imagine that BRCA1 and BRCA2 may also be involved in host-virus interactions, much like the story of Nbs1 and ICP0. Many of the known functions of BRCA1, which include a role in transcription, DNA repair, cell cycle regulation, and centrosome maintenance, could feasibly play a role in viral lifecycles.

Interestingly, there is increasing evidence to suggest that BRCA1 plays a critical role in the developing brain (119). Conditional knockout of *BRCA1* in mice results in high levels of apoptosis and an abnormally small brain size. In addition, expression of

BRCA1 is regulated by microcephalin, a protein involved in DNA repair and cell cycle regulation (120). Mutations in the microcephalin gene, *MCPH1*, result in a disorder characterized by a small head circumference and a decreased size of the brain. Based on these findings, it is possible that selection is acting upon BRCA1 functions that are critical during brain development. It will be interesting to see if the sites of positive selection we have identified have an effect on this important embryological process and if this effect correlates to the known brain masses of primate species used in our study.

NBS1 AND HSV-1

The cellular response to DNA damage is a highly coordinated and complex process. Several mechanisms exist to safeguard against the many types of lesions that can be incurred, ultimately resulting in the repair of damage, senescence, or apoptosis. One of the most deleterious types of DNA damage is the double-strand break, which is recognized by the Mre11-Rad50-Nbs1 (MRN complex) (121). ATM rapidly localizes to sites of damage via interactions with MRN, leading to autophosphorylation and subsequent activation. Once activated, ATM mediates the phosphorylation of several other repair factors such as H2AX, MDC1, and Nbs1, resulting in the retention of these proteins at the break. Recruitment of the E3 ubiquitin ligases, RNF8 and RNF168, to the sites of damage results in several additional post-translation modification events that ultimately leads to the accumulation of repair factors and chromatin remodeling near the sites of damage. Additionally, it has become increasingly evident that viruses that replicate within the nucleus activate these same pathways and can be considered as yet another form of genotoxic stress. In chapter 4, I describe a way in which HSV-1 has

evolved to utilize the functions of the DNA repair protein, Nbs1, to enhance the viral lifecycle.

When HSV-1 enters the cell, the intact capsid core travels to the nucleus where it docks onto the nuclear pore complex. The ~150 kb genome is confined within a relatively small 125 nm icosahedral shell, resulting in the ejection of viral contents into the nucleus through a pressure-driven mechanism (122). Incoming viral genomes rapidly co-localize with DNA repair factors such as γH2AX and MDC1 at sites of viral entry (68). In fact, many of these host factors are the same proteins that mark sites of double-strand breaks in host chromosomal DNA. In the absence of the virally encoded ICP0 protein, the recruitment of RNF8 and RNF168 results in the deposition of repressive marks onto the viral genome that suppress viral transcription. In order to subvert this defense strategy employed by the host, the ICP0 protein of HSV-1 ubiquitinates and degrades the antiviral effects mediated by RNF8 and RNF168 (94). Therefore, the DNA damage response during HSV-1 entry into the nucleus can be viewed as an intrinsic defense mechanism against viral infection.

However, other components of the ATM signaling pathway have been shown to greatly enhance the viral lifecycle. For example, Mre11- and ATM-deficient cells exhibit much lower yields in virus production (65, 91). In addition, phosphorylation of the checkpoint kinase, chk2, by ATM is required for inducing G2/M arrest for maximal viral replication (92). Collectively, this suggests that HSV-1 may be inducing the ATM signaling pathway in order to induce cell cycle arrest, while inhibiting the aspect of ATM signaling that results in transcriptional repression of viral genomes. However, initial activation event leading to ATM signaling is still unknown.

In our study, we show that human Nbs1 also enhances the viral lifecycle, much like the previously reported effect of Mre11, ATM, and chk2. In addition, we identify a

naturally occurring Nbs1 variant from a non-human primate that is not susceptible to hijacking by the virus. We show that variants of Nbs1 that support HSV-1 replication interact with ICP0, while those that do not enhance virus production are not able to associate with ICP0 (data not shown).

The results from our study suggest that the successful interaction between Nbs1 and ICP0 results in the ubiquitination of Nbs1 and this may be the reason why Nbs1 exerts a positive effect on virus replication. However, how these events specifically contribute to the HSV-1 lifecycle is still undetermined. Interestingly, Nbs1 has been shown to be ubiquitinated by the cellular E3 ubiquitin ligase skp2 in response to DNA damage, allowing for the binding and activation of ATM (123). One can imagine that ubiquitination of Nbs1 by ICP0 could be mimicking this cellular process, resulting in the activation of the DNA damage response. In fact, expression of ICP0 alone has been demonstrated to induce autophosphorylation and activation of ATM (92). If this were true, then Nbs1 may be the first contact point in which ICP0 activates the ATM pathway. To our knowledge, this is the first time in which a model has been presented that specifically addresses how the ATM pathway is initially activated during the early stages of the HSV-1 lifecycle. This model can also explain why some interactions between ICP0 and DNA repair factors or checkpoint regulators do not result in degradation of the host proteins and ascribes a new function for ICP0 in both activating and disarming certain aspects of this major signaling pathway. Additional experiments are currently underway to test whether this is indeed the case. It will be interesting to see if there are any differences in ATM activation between our cell lines during infection with the wild-type and ICP0-null viruses.

In addition to the mechanistic insights we have uncovered in our study of the HSV-1 lifecycle, we show that ICP0 interacts with Nbs1 in a species-specific manner and

that this interaction correlates strongly to an increase in viral titers. Using naturally occurring primate alleles, we identified four residues critical for this interaction. We also show for the first time that a protein that is neither a restriction factor or a cellular entry receptor may act as a barrier to cross-species transmission of HSV-1 and that the ability of specific primate Nbs1 variants to evade interaction with ICP0 limits viral replication. This phenomenon may be a result of long-term coevolution between primate hosts and their cognate herpes simplex viruses. In this scenario, the ICP0 of a specific herpes simplex virus may have evolved the ability to interact with the Nbs1 of its primate host in order to utilize its functions for the advancing the viral lifecycle. However, when this ICP0 variant encounters a new Nbs1 from a non-host species, subtle sequence differences within the ICP0 binding region of Nbs1 may alter the ability of this particular ICP0 variant to establish an interaction. Moreover, this variability in Nbs1 of the non-host species may be a result of coevolution with the ICP0 of its own herpes simplex virus. It will be interesting to see if other ICP0 proteins encoded by primate herpes simplex viruses exhibit a different pattern of compatibility with primate Nbs1 proteins or whether this inability of the white-cheeked gibbon Nbs1 to evade ICP0 interaction is unique evolutionary accident. Attempts to experimentally evolve HSV-1 in cells expressing the white-cheeked gibbon Nbs1 were not successful, suggesting that the latter scenario may be the case. However, since HSV-1 is a considerably slow evolving virus, extensive serial passaging of the virus may be required. Nonetheless, we demonstrate the power in using primate alleles for mapping the interface of dynamic host-virus interactions.

Appendix A

CIRCLE SEQUENCING BIOCHEMICAL PROTOCOL

Materials:

DNA of interest

Tris-EDTA

Low Molecular Weight Ladder (NEB N3233L)

Low melting point agarose (NuSieve GTG Lonza 50081)

TBE

SYBR Gold (Invitrogen S-11494)

QIAEX II Gel Extraction Kit (QIAGEN 20021)

DNase-, RNase-free water

T4 PNK (NEB M0201S)

10 mM ATP (NEB P0756S)

QIAGEN MinElute Purification Kit (QIAGEN 28004)

Liquid nitrogen

CircLigase II (Epicentre CL9025K)

Exonuclease I (NEB M0293S)

Exonuclease III (NEB M0206S)

2X Annealing Buffer (10 mM Tris pH8, 50 mM NaCl, 1 mM EDTA)

Exo-Resistant Random Primers (Thermo SO181)

Phi29 DNA Polymerase (NEB M0269L)

10 mM dNTP (NEB N0447L)

Inorganic Pyrophosphatase (NEB M0361S)

Uracil-DNA Glycosylase (NEB M0280S)

Formamidopyrimidine-DNA Glycosylase (NEB M0240S)

3M Sodium Acetate pH 5.2

100% Ethanol

70% Ethanol

MiSeq library preparation kit

Ampure XP beads (Agencourt A63880)

Magnetic stand

KAPA HiFi DNA polymerase (KAPA Biosystems KK2101)

Covaris S220

Microtubes for shearing

Nanodrop

Thermocycler or water bath

Heat block

MiSeq

Procedure:

A. DNA shearing (Begin here if sequencing genomic DNA)

1. Shear gDNA resuspended in TE to 150 bp using Covaris S220. Shear about 10 μg at a time in 130 μl total volume.

   Duty cycle – 10%

   Intensity – 5

   Cycles per burst – 200

   Time – 14 min

   Note: Shearing conditions will vary depending on machine. Also, it is important that the DNA be resuspended in TE. DNA in H_2O gives a broader length distribution for unknown reasons. Also, use Invitrogen's DNA quant

machine to quantitate yeast DNA preps. We have noticed that contaminants in the prep give inaccurate readings on the spec. Run 10 μl of sheared product out on a 1.5% gel to determine if shearing was efficient. If needed, combine tubes and use Speedvac to concentrate.

2. Run at least 2.5 μg of sheared DNA in 1 lane and a low MW DNA ladder in a separate lane of a 1.5% low melting point agarose gel in 1X TBE. Cut off the lane with the DNA ladder and stain with SYBR Gold (1:10,000 dilution in TBE). Place next to the unstained remainder of the gel. Use that placement as a guide to cut 150 bp fragments from the sheared gDNA sample lane.

   Optional: Stain the unused portion of the gel to see if right size fragments were cut from the gel. It is advisable to also cut slices slightly higher (175 bp, 200 bp, 250 bp) and run these fragments on a Bioanalyzer to ensure that fragment size is centered around 150 bp.

   Note: The average size of the fragments is important. They should be slightly less than 1/3 of the average read length of the sequencing technology that you are using. This will ensure that you get at least 3 repeats of your sequence in 1 read length.

3. Gel extract the excised portion of material using the QIAEX II Gel Extraction Kit. Determine concentration using nanodrop.

B. DNA modification and preparation (Begin here if sequencing amplicons)

4. Phosphorylate the gel extracted fragments using T4 PNK.

5 μl	10X T4 PNK buffer
5 μl	10 mM ATP
$\leq$ 300 pmol	sheared DNA

to 49 μl	H2O
1 μl	T4 PNK (10U)

Incubate at 37°C for 30 min.

5. Purify DNA using QIAGEN MinElute Purification Kit. Elute in 20 μl of H_2O.

   Note: This kit claims to purify PCR products 70 bp to 4 kb

6. Denature DNA by incubating at 95°C for 15 min, preferably in a heat block. Immediately remove from heat block and snap freeze in liquid nitrogen for 5 min. Place frozen sample on ice and allow to melt slowly.

   Note: This is a critical step in which single-stranded DNA is generated for the circularization reaction. Freezing in liquid nitrogen ensures that rapid cooling occurs, preventing reannealing of DNA into double-stranded fragments.

7. Determine concentration of DNA using nanodrop. Make sure to use ssDNA settings when determining concentration.

C. Circularization

8. Set up circularization reactions.

2 μl	CircLigase II 10X Reaction Buffer
1 μl	50 mM $MnCl_2$
5 pmol	ssDNA (~240 ng for 150 bases)
to 19 μl	H_2O
1 μl	CircLigase II ssDNA Ligase (100U)

   Incubate at 60°C for 1-16 hours.

   Note: It is convenient to set up this reaction at the end of the day and let it go overnight.

9. Perform a 2nd round of circularization by adding 1 μl of CircLigase II to the reaction and incubating at 60°C for an additional 6 hours. Heat inactivate the enzyme at 80°C for 10 min.

   Note: The addition of more enzyme is not necessary but can increase the overall yield of circularized product in some instances.

10. Digest any linear DNA fragments still remaining in the reaction by adding 1 μl Exonuclease I (20U) and 0.5 μl Exonuclease III (50U). Incubate at 37°C for 1 hour and heat inactivate at 80°C for 10 min.

11. Purify the circularized products using QIAGEN MinElute clean up kit and elute in 20 μl H_2O. Determine concentration using ssDNA settings.

    Note: You should recover about 10-20% of your input DNA.

D. Rolling Circle Amplification (RCA)

12. Anneal random primers to DNA circles.

10 μl	2X Annealing Buffer
	(10 mM Tris pH8, 50 mM NaCl, 1 mM EDTA)
1 μl	Exo-Resistant Random Primers
1-100 ng	circularized DNA
to 20 μl	H_2O

    Incubate at 95°C for 5 min and cool to 4°C.

    Note: This step is performed in a thermocycler. The ramp rate of cooling is not crucial for primer annealing.

13. Set up RCA reaction on ice.

5 μl	10X Phi29 DNA Polymerase Reaction Buffer
1 μl	100X BSA

1 μl	10 mM dNTP
20 μl	primer-annealed circularized DNA
17 μl	H_2O
1 μl	Inorganic Pyrophosphatase (0.2 U)
1 μl	Uracil-DNA Glycosylase (10U)
1 μl	Formamidopyrimidine-DNA Glycosylase (16U)
2 μl	phi29 DNA Polymerase

Incubate at 30ºC for 3 hours. Heat inactivate at 65ºC for 10 min.

Note: Two modifications have been made to the traditional Phi29 reaction. Addition of Uracil-DNA Glycosylase removes any deaminated cytosines from DNA. Formamidopyrimidine-DNA Glycosylase is also added to remove 8-oxo-guanine products. Both of these types of damaged bases will result in errors in the DNA sequence in downstream processes.

Note: If less than 10 ng of circles are used in for RCA, longer amplification times (more than 3 hours) may be required to obtain sufficient amounts of product for library preparation.

14. Purify RCA products by ethanol precipitation. Resuspend DNA in TE.

E. Library preparation for NGS

Note: The remainder of the protocol just describes standard Illumina MiSeq library preparation. The only modifications to the manufacturer-recommended protocol involve the shearing of RCA products and the purification of final library preps. The following protocol produces libraries for MiSeq 2x250 reads.

Sample preparation:

15. Shear RCA products to 1,500 bp using Covaris S220.

Duty cycle – 2%

Intensity – 4

Cycles per burst – 200

Time – 15 sec

16. Purify DNA using MinElute Purification Kit. Elute in 20 μl H_2O.

17. Prepare adaptor ligations as suggested by manufacturer. (EX. TruSeq)

18. Size select adaptor-ligated samples using Ampure XP beads.

    Bring Ampure XP beads to room temperature (about 30 min).

    Add 0.6X volume of beads to sample and vortex for 1 min. Spin briefly.

    Incubate at room temperature for 15 min.

    Magnetize sample for 5 min.

    Remove supernatant and discard.

    Wash with fresh 70% ethanol two times.

    Dry beads for 5 min.

    Add 70 μl of TE to beads and vortex for 1 min. Spin briefly and incubate at room temperature for 5 min.

    Magnetize sample for 5 min.

    Remove supernatant and place in new tube.

19. Set up PCRs to amplify library.

10 μl	5X HiFi Buffer
1.5 μl	10 mM dNTP
1.5 μl	10 uM P5 PCR primer
1.5 μl	10 uM P7 PCR primer
34.5 μl	DNA
1 μl	KAPA HiFi

95ºC	2 min	
98C	20 sec	
60C	15 sec	10 cycles
72ºC	2 min	
72ºC	5 min	
4ºC	hold	

20. Size select PCR products using 0.6X volume Ampure XP beads (Step 18). Alternatively, gel extract 1.2 kb fragments with 1% low melting point agarose gel using QIAEX II kit.

References

1. Loman NJ et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* 30:434–439.

2. Jünemann S et al. (2013) Updating benchtop sequencing performance comparison. *Nature Biotechnology* 31:294–296.

3. Liu L et al. (2012) Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* 2012:1–11.

4. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci USA* 108:20166–20171.

5. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530–9535.

6. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research* 39:e81.

7. Schmitt MW et al. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci USA* 109:14508–14513.

8. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J (2010) Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Meth* 7:119–122.

9. Jackson SP, Bartek J (2009) The DNA-damage response in human biology and disease. *Nature* 461:1071–1078.

10. Iyama T, Wilson DM (2013) DNA repair mechanisms in dividing and non-dividing cells. *DNA Repair (Amst)* 12:620–636.

11. Demogines A et al. (2010) Ancient and Recent Adaptive Evolution of Primate Non-Homologous End Joining Genes. *PLoS Genet* 6:e1001169.

12. O'Connell MJ (2010) Selection and the cell cycle: positive Darwinian selection in a well-known DNA damage response pathway. *J Mol Evol* 71:444–457.

13. Huttley GA et al. (2000) Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. *Nat Genet* 25:410–413.

14. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19:908–917.

15. Fleming MA, Potter JD, Ramirez CJ, Ostrander GK, Ostrander EA (2003) Understanding missense mutations in the BRCA1 gene: an evolutionary approach. *Proc Natl Acad Sci U S A* 100:1151–1156.

16. Burk-Herrick A, Scally M, Amrine-Madsen H, Stanhope MJ, Springer MS (2006) Natural selection and mammalian BRCA1 sequences: elucidating functionally important sites relevant to breast cancer susceptibility in humans - Springer. *Mamm Genome* 17:257–270. Available at: http://link.springer.com/article/10.1007%2Fs00335-005-0067-2.

17. Pavlicek A et al. (2004) Evolution of the tumor suppressor BRCA1 locus in primates: implications for cancer predisposition. *Hum Mol Genet* 13:2737–2751.

18. Albert B Sabin AMW (1934) Acute ascending myelitis following a monkey bite, with the isolation of a virus capable of reproducing the disease. *The Journal of Experimental Medicine* 59:115.

19. Gay FP, Holden M (1933) The Herpes Encephalitis Problem, II. *The Journal of Infectious Diseases* 53:287–303.

20. Jennifer L Huff PAB (2003) B-Virus (Cercopithecine herpesvirus 1) Infection in Humans and Macaques: Potential for Zoonotic Disease. *Emerg Infect Dis* 9:246.

21. Parrish CR et al. (2008) Cross-Species Virus Transmission and the Emergence of New Epidemic Diseases. *Microbiology and Molecular Biology Reviews* 72:457–470.

22. Sawyer SL, Elde NC (2012) A cross-species view on viruses. *Curr Opin Virol* 2:561–568.

23. Holmes EC (2013) What can we predict about viral evolution and emergence? *Curr Opin Virol* 3:180–184.

24. Meacham F et al. (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 12:451.

25. Reumers J et al. (2012) Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nature Biotechnology* 30:61–68.

26. Roach JC et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639.

27. Lou DI, Hussmann JA et al. (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci USA* 110:19872–19877.

28. Acevedo A, Brodsky L, Andino R (2014) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* 505:686–690.

29. Eid J et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.

30. Mullen P et al. (1997) BRCA1 5382insC mutation in sporadic and familial breast and ovarian carcinoma in Scotland. *Br J Cancer* 75:1377–1380.

31. O'Donovan PJ, Livingston DM (2010) BRCA1 and BRCA2: breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair. *Carcinogenesis* 31:961–967.

32. Hemel D, Domchek SM (2010) Breast Cancer Predisposition Syndromes. *Hematology/Oncology Clinics of North America* 24:799–814.

33. Ludwig T, Chapman DL, Papaioannou VE, Efstratiadis A (1997) Targeted mutations of breast cancer susceptibility gene homologs in mice: lethal phenotypes of Brca1, Brca2, Brca1/Brca2, Brca1/p53, and Brca2/p53 nullizygous embryos. *Genes Dev* 11:1226–1241.

34. Hurst LD, Pál C (2001) Evidence for purifying selection acting on silent sites in BRCA1. *Trends Genet* 17:62–65.

35. Larkin MA et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.

36. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556.

37. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15:568–573.

38. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11:725–736.

39. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

40. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–

936.

41. Wong WSW, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051.

42. Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118.

43. Rodriguez S, Gaunt TR, Day INM (2009) Hardy-Weinberg equilibrium testing of biological ascertainment for Mendelian randomization studies. *Am J Epidemiol* 169:505–514.

44. Meyerson NR, Sawyer SL (2011) Two-stepping through time: mammals and viruses. *Trends Microbiol* 19:286–294.

45. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.

46. Durocher F et al. (1996) Comparison of BRCA1 polymorphisms, rare sequence variants and/or missense mutations in unaffected and breast/ovarian cancer populations. *Hum Mol Genet* 5:835–842.

47. Dunning AM et al. (1997) Common BRCA1 variants and susceptibility to breast and ovarian cancer in the general population. *Hum Mol Genet* 6:285–289.

48. Mizuta R et al. (1997) RAB22 and RAB163/mouse BRCA2: proteins that specifically interact with the RAD51 protein. *Proc Natl Acad Sci U S A* 94:6927–6932.

49. Wong AKC, Pero R, Ormonde PA, Tavtigian SV, Bartel PL (1997) RAD51 interacts with the evolutionarily conserved BRC motifs in the human breast cancer susceptibility gene brca2. *Journal of Biological Chemistry* 272:31941–31944.

50. Holloman WK (2011) Unraveling the mechanism of BRCA2 in homologous recombination. *Nat Struct Mol Biol* 18:748–754.

51. Rajendra E, Venkitaraman AR (2009) Two modules in the BRC repeats of BRCA2 mediate structural and functional interactions with the RAD51 recombinase. *Nucleic Acids Research* 38:82–96.

52. Pellegrini L et al. (2002) Insights into DNA recombination from the structure of a RAD51-BRCA2 complex. *Nature* 420:287–293.

53. Clark AG et al. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302:1960–1963.

54. Vallender EJ, Lahn BT (2004) Positive selection on the human genome. *Hum Mol Genet* 13 Review Issue 2:R245–54.

55. Sawyer SL, Emerman M, Malik HS (2004) Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol* 2:E275.

56. Sawyer SL, Wu LI, Emerman M, Malik HS (2005) Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proc Natl Acad Sci U S A* 102:2832–2837.

57. Elde NC, Child SJ, Geballe AP, Malik HS (2009) Protein kinase R reveals an evolutionary model for defeating viral mimicry. *Nature* 457:485–490.

58. Lim ES, Malik HS, Emerman M (2010) Ancient Adaptive Evolution of Tetherin Shaped the Functions of Vpu and Nef in Human Immunodeficiency Virus and Primate Lentiviruses. *J Virol* 84:7124–7134.

59. Laguette N et al. (2012) Evolutionary and Functional Analyses of the Interaction between the Myeloid Restriction Factor SAMHD1 and the Lentiviral Vpx Protein. *Cell Host and Microbe* 11:205–217.

60. Lim ES et al. (2012) The Ability of Primate Lentiviruses to Degrade the Monocyte Restriction Factor SAMHD1 Preceded the Birth of the Viral Accessory Protein Vpx. *Cell Host and Microbe* 11:194–204.

61. Sawyer SL, Malik HS (2006) Positive selection of yeast nonhomologous end-joining genes and a retrotransposon conflict hypothesis. *Proc Natl Acad Sci U S A* 103:17614–17619.

62. Lilley CE, Schwartz RA, Weitzman MD (2007) Using or abusing: viruses and the cellular DNA damage response. *Trends Microbiol* 15:119–126.

63. Chaurushiya MS, Weitzman MD (2009) Viral manipulation of DNA repair and cell cycle checkpoints. *DNA Repair (Amst)* 8:1166–1176.

64. Stracker TH, Carson CT, Weitzman MD (2002) Adenovirus oncoproteins inactivate the Mre11-Rad50-NBS1 DNA repair complex. *Nature* 418:348–352.

65. Lilley CE, Carson CT, Muotri AR, Gage FH, Weitzman MD (2005) DNA repair proteins affect the lifecycle of herpes simplex virus 1. *Proc Natl Acad Sci USA* 102:5844–5849.

66. Mohni KN, mastrocola AS, bai P, Weller SK, heinen CD (2011) DNA mismatch repair proteins are required for efficient herpes simplex virus 1 replication. *J Virol* 85:12241–12253.

67. Lees-Miller SP et al. (1996) Attenuation of DNA-dependent protein kinase activity and its catalytic subunit by the herpes simplex virus type 1 transactivator ICP0. *The Journal of Virology* 70:7471–7477.

68. Lilley CE, Chaurushiya MS, Boutell C, Everett RD, Weitzman MD (2011) The intrinsic antiviral defense to incoming HSV-1 genomes includes specific DNA repair proteins and is counteracted by the viral protein ICP0. *PLoS Pathog* 7:e1002084.

69. Zimmerman ES et al. (2004) Human immunodeficiency virus type 1 Vpr-mediated G2 arrest requires Rad17 and Hus1 and induces nuclear BRCA1 and gamma-H2AX focus formation. *Mol Cell Biol* 24:9286–9294.

70. Nakai-Murakami C et al. (2006) HIV-1 Vpr induces ATM-dependent cellular signal with enhanced homologous recombination. *Oncogene* 26:477–486.

71. Daniel R, Katz RA, Skalka AM (1999) A Role for DNA-PK in Retroviral DNA Integration. *Science* 284:644–647.

72. Daniel R et al. (2004) Evidence that stable retroviral transduction and cell survival following DNA integration depend on components of the nonhomologous end joining repair pathway. *The Journal of Virology* 78:8573–8581.

73. Smith JA et al. (2008) Evidence that the Nijmegen breakage syndrome protein, an early sensor of double-strand DNA breaks (DSB), is involved in HIV-1 post-integration repair by recruiting the ataxia telangiectasia-mutated kinase in a process similar to, but distinct from, cellular DSB repair. *Virol J* 5:11.

74. Zhong Q, Chen C-F, Chen P-L, Lee W-H (2002) BRCA1 Facilitates Microhomology-mediated End Joining of DNA Double Strand Breaks. *Journal of Biological Chemistry* 277:28641–28647.

75. Lau A, Kanaar R, Jackson SP, O'Connor MJ (2004) Suppression of retroviral infection by the RAD52 DNA repair protein. *EMBO J* 23:3421–3429.

76. Lloyd AG et al. (2006) Effect of DNA Repair Protein Rad18 on Viral Infection. *PLoS Pathog* 2:e40.

77. Cosnefroy O et al. (2011) Stimulation of hRAD51 nucleofilament restricts HIV-1

integration in vitro and in infected cells. *J Virol*.

78. Carter AJ, Nguyen AQ (2011) Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. *BMC Med Genet* 12:160.

79. Crespi BJ, Summers K (2006) Positive selection in the evolution of cancer. *Biol Rev* 81:407.

80. Chayavichitsilp P, Buckwalter JV, Krakowski AC, Friedlander SF (2009) Herpes Simplex. *Pediatrics in Review* 30:119–130.

81. McGeoch DJ, Cook S (1994) Molecular phylogeny of the alphaherpesvirinae subfamily and a proposed evolutionary timescale. *J Mol Biol* 238:9–22.

82. McGeoch DJ, Cook S, Dolan A, Jamieson FE, Telford EA (1995) Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *J Mol Biol* 247:443–458.

83. McGeoch DJ, Dolan A, Ralph AC (2000) Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *The Journal of Virology* 74:10401–10406.

84. Wertheim JO, Smith MD, Smith DM, Scheffler K, Kosakovsky Pond SL (2014) Evolutionary Origins of Human Herpes Simplex Viruses 1 and 2. *Mol Biol Evol*.

85. Cohen JI et al. (2002) Recommendations for Prevention of and Therapy for Exposure to B Virus (Cercopithecine Herpesvirus1). *Clinical Infectious Diseases* 35:1191–1203.

86. Matz-Rensing K et al. (2003) Fatal Herpes simplex Infection in a Group of Common Marmosets (Callithrix jacchus). *Veterinary Pathology* 40:405–411.

87. Landolfi JA, Wellehan JFX, Johnson AJ, Kinsel MJ (2005) Fatal human herpesvirus type 1 infection in a white-handed gibbon (Hylobates lar). *J Vet Diagn Invest* 17:369–371.

88. Huemer HP, Larcher C, Czedik-Eysenberg T, Nowotny N, Reifinger M (2002) Fatal Infection of a Pet Monkey with Human herpesvirus 1. *Emerg Infect Dis* 8:639–641.

89. Rhesus Macaque Genome Sequencing and Analysis Consortium et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222–234.

90. Wilkinson DE, Weller SK (2004) Recruitment of cellular recombination and repair proteins to sites of herpes simplex virus type 1 DNA replication is dependent on the composition of viral proteins within prereplicative sites and correlates with the induction of the DNA damage response. *The Journal of Virology* 78:4783–4796.

91. Shirata N et al. (2005) Activation of Ataxia Telangiectasia-mutated DNA Damage Checkpoint Signal Transduction Elicited by Herpes Simplex Virus Infection. *Journal of Biological Chemistry* 280:30336–30341.

92. Li H et al. (2008) Chk2 is required for HSV-1 ICP0-mediated G2/M arrest and enhancement of virus growth. *Virology* 375:13–23.

93. Parkinson J, Lees-Miller SP, Everett RD (1999) Herpes simplex virus type 1 immediate-early protein vmw110 induces the proteasome-dependent degradation of the catalytic subunit of DNA-dependent protein kinase. *The Journal of Virology* 73:650–657.

94. Lilley CE et al. (2010) A viral E3 ligase targets RNF8 and RNF168 to control histone ubiquitination and DNA damage responses. *The EMBO Journal* 29:943–955.

95. Wilkinson DE, Weller SK (2006) Herpes simplex virus type I disrupts the ATR-dependent DNA-damage response during lytic infection. *J Cell Sci* 119:2695–2703.

96. Mohni KN, Smith S, Dee AR, Schumacher AJ, Weller SK (2013) Herpes simplex virus type 1 single strand DNA binding protein and helicase/primase complex disable cellular ATR signaling. *PLoS Pathog* 9:e1003652.

97. Chaurushiya MS et al. (2012) Viral E3 Ubiquitin Ligase-Mediated Degradation of a Cellular E3: Viral Mimicry of a Cellular Phosphorylation Mark Targets the RNF8 FHA Domain. *Mol Cell* 46:79–90.

98. Taylor TJ, Knipe DM (2004) Proteomics of herpes simplex virus replication compartments: association of cellular DNA replication, repair, recombination, and chromatin remodeling proteins with ICP8. *The Journal of Virology* 78:5856–5866.

99. Balasubramanian N, Bai P, Buchek G, Korza G, Weller SK (2010) Physical Interaction between the Herpes Simplex Virus Type 1 Exonuclease, UL12, and the DNA Double-Strand Break-Sensing MRN Complex. *J Virol* 84:12504–12514.

100. Lou DI et al. (2014) Rapid evolution of BRCA1 and BRCA2 in humans and other primates. *BMC Evol Biol* 14:155.

101. Stow ND, Stow EC (1986) Isolation and characterization of a herpes simplex virus type 1 mutant containing a deletion within the gene encoding the immediate early polypeptide Vmw110. *J Gen Virol* 67:2571–2585.

102. Everett RD (1989) Construction and characterization of herpes simplex virus type 1 mutants with defined lesions in immediate early gene 1. *J Gen Virol* 70:1185–1202.

103. Kraakman-van der Zwet M et al. (1999) Immortalization and characterization of Nijmegen Breakage Syndrome fibroblasts. *Mutation Research/DNA Repair* 434:17–27.

104. Lloyd J et al. (2009) A Supramodular FHA/BRCT-Repeat Architecture Mediates Nbs1 Adaptor Function in Response to DNA Damage. *Cell* 139:100–111.

105. Williams RS et al. (2009) Nbs1 Flexibly Tethers Ctp1 and Mre11-Rad50 to Coordinate DNA Double-Strand Break Processing and Repair. *Cell* 139:87–99.

106. Schiller CB et al. (2012) Structure of Mre11–Nbs1 complex yields insights into ataxia-telangiectasia–like disease mutations and DNA damage signaling. *Nat Struct Mol Biol* 19:693–700.

107. Falck J, Coates J, Jackson SP (2005) Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. *Nature* 434:605–611.

108. Stracker TH et al. (2005) Serotype-specific reorganization of the Mre11 complex by adenoviral E4orf3 proteins. *J Virol* 79:6664–6673.

109. Araujo FD, Stracker TH, Carson CT, Lee DV, Weitzman MD (2005) Adenovirus type 5 E4orf3 protein targets the Mre11 complex to cytoplasmic aggresomes. *J Virol* 79:11382–11391.

110. Forrester NA et al. (2011) Serotype-specific inactivation of the cellular DNA damage response during adenovirus infection. *J Virol* 85:2201–2211.

111. Evans JD, Hearing P (2005) Relocalization of the Mre11-Rad50-Nbs1 complex by the adenovirus E4 ORF3 protein is required for viral replication. *J Virol* 79:6207–6215.

112. Baker A, Rohleder KJ, Hanakahi LA, Ketner G (2007) Adenovirus E4 34k and E1b 55k oncoproteins target host DNA ligase IV for proteasomal degradation. *The Journal of Virology* 81:7034–7040.

113. Orazio NI, Naeger CM, Karlseder J, Weitzman MD (2011) The Adenovirus E1b55K/E4orf6 Complex Induces Degradation of the Bloom Helicase during Infection. *J Virol* 85:1887–1892.

114. Blackford AN et al. (2010) Adenovirus 12 E4orf6 inhibits ATR activation by promoting TOPBP1 degradation. *Proc Natl Acad Sci USA* 107:12251–12256.

115. Israfil H, Zehr SM, Mootnick AR, Ruvolo M, Steiper ME (2011) Unresolved molecular phylogenies of gibbons and siamangs (Family: Hylobatidae) based on mitochondrial, Y-linked, and X-linked loci indicate a rapid Miocene radiation or sudden vicariance event. *Molecular Phylogenetics and Evolution* 58:447–455.

116. Chelbi-Alix MK, de The H (1999) Herpes virus induced proteasome-dependent degradation of the nuclear bodies-associated PML and Sp100 proteins. *Oncogene* 18:935–941.

117. Lomonte P, Morency E (2007) Centromeric protein CENP-B proteasomal degradation induced by the viral protein ICP0. *FEBS Lett* 581:658–662.

118. Orzalli MH, DeLuca NA, Knipe DM (2012) Nuclear IFI16 induction of IRF-3 signaling during herpesviral infection and degradation of IFI16 by the viral ICP0 protein. *Proc Natl Acad Sci USA* 109:E3008–17.

119. Pao GM et al. (2014) Role of BRCA1 in brain development. *Proc Natl Acad Sci USA* 111:E1240–E1248.

120. Yang S-Z, Lin F-T, Lin W-C (2008) MCPH1/BRIT1 cooperates with E2F1 in the activation of checkpoint, DNA repair and apoptosis. *EMBO Rep* 9:907–915.

121. Cristina Bartocci ELD (2013) Put a RING on it: regulation and inhibition of RNF8 and RNF168 RING finger E3 ligases at DNA damage sites. *Frontiers in Genetics* 4.

122. Bauer DW, Huffman JB, Homa FL, Evilevitch A (2013) Herpes virus genome, the pressure is on. *J Am Chem Soc* 135:11216–11221.

123. Wu J et al. (2012) Skp2 E3 Ligase Integrates ATM Activation and Homologous Recombination Repair by Ubiquitinating NBS1. *Mol Cell* 46:351–361.